© 2000 Oxford University Press Nucleic Acids Research, 2000, Vol. 28, No. 1 123–125

APPENDIX A

WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction

Ross Overbeek[1,*], Niels Larsen[1], Gordon D. Pusch[1], Mark D'Souza[1,2], Evgeni Selkov Jr[1,2], Nikos Kyrpides[1], Michael Fonstein[1], Natalia Maltsev[2] and Evgeni Selkov[1,2]

[1]Integrated Genomics Inc., 2201 W. Campbell Park Drive, Chicago, IL 60612, USA and [2]Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439, USA

Received September 1, 1999; Revised and Accepted October 13, 1999

ABSTRACT

The WIT (What Is There) (http://wit.mcs.anl.gov/WIT2/) system has been designed to support comparative analysis of sequenced genomes and to generate metabolic reconstructions based on chromosomal sequences and metabolic modules from the EMP/MPW family of databases. This system contains data derived from about 40 completed or nearly completed genomes. Sequence homologies, various ORF-clustering algorithms, relative gene positions on the chromosome and placement of gene products in metabolic pathways (metabolic reconstruction) can be used for the assignment of gene functions and for development of overviews of genomes within WIT. The integration of a large number of phylogenetically diverse genomes in WIT facilitates the understanding of the physiology of different organisms.

INTRODUCTION

Starting with *Haemophilus influenza* (1) in 1995, over 20 microbial organisms have had their total genomic DNA sequenced and almost 100 others have been started as shown in the GOLD database (2). Currently we are observing an impressive development of the human genome project (3,4). In response to this growing amount of sequence data, computational tools for genome analysis have been developed and merged into shared analytical environments, such as GeneQuiz (5), KEGG (6), Pedant (7) and Entrez Genomes (8), moving cross-genome analysis to a new level. The development of analytical systems, together with the growth of sequencing data, have increased gene recognition rates from <50% (9,10) to >70% (11,12). Today, this remaining 30%, so-called 'hypothetical' or 'orphan' genes, separates us from a complete description of the genomic content and functions of an organism.

Computational approaches based on various types of clustering of potential genes, whether in phylogenetic space, as clusters of orthologous genes (COGs) (13) or position on the chromosome, such as in operons (14), increase the gene assignment level even further. An important stage of genome analysis is the integration of gene assignments into an organism-specific overview via so-called functional reconstruction (15), which is the conceptual assembly of metabolic pathways, transport units and signal transduction pathways. It allows reconciliation of inconsistencies between different types of analysis, and often results in changes of initial gene function assignments based on similarity scoring.

The WIT system, discussed in this paper, represents the development of a genome analysis strategy in a multi-genome environment, which combines a variety of tools, dealing with individual open reading frames (ORFs) or proteins, with the ability to derive general conclusions. Using the WIT genome analysis system, a major part of the central metabolism of an organism can be reconstructed entirely *in silico* (16).

WIT: A VIEW TO A GENOME

The current version of the WIT system is available at Argonne National Laboratory (http://wit.mcs.anl.gov/WIT2) or at Integrated Genomics Inc. (http://wit.IntegratedGenomics.com/IGwit) and contains 43 complete or nearly complete genomes (Table 1).

These genomes consist of 123 482 predicted ORFs, of which 78 144 could be given functional assignments and 41 742 could be assembled into metabolic pathways, which came from EMP/MPW database (15). Pathways involved in the metabolism of carbohydrates and amino acids are connected into schematic overviews allowing the user to reveal substrates and final products connecting metabolic modules.

In order to incorporate a genome into WIT, a gene-searching program called CRITICA (17) can be used. Potential coding regions recognized in the DNA contigs are subjected to a FASTA search against the non-redundant database of assigned genes and loaded into the WIT system, together with the pre-computed tables of best hits.

WIT provides a set of tools for the characterization of gene structures and functions, such as Functional Coupling, or Preserved Operons. WIT also provides integrated WWW access to such tools as PSI-BLAST, PROSITE, ProDom, COG, ClustalW and others. Functional content may be queried, for example, by looking for specific functions missing in the metabolic pathways, or by separating alternative gene functions derived from similarities found for a putative gene.

After genes have been assigned initial functions, they are then 'attached' to pathways by choosing templates from metabolic database (MPW) which best incorporate all observed functions. For any given organism, this usually leads to identification of

[*]To whom correspondence should be addressed. Tel: +1 312 491 0846; Fax: +1 312 491 0856; Email: ross@integratedgenomics.com

Table 1. Genomes in WIT

Eukarya	*Saccharomyces cerevisiae, Caenorhabditis elegans*
Archaea	*Sulfolobus solfataricus, Archaeoglobus fulgidus,Halobacterium* sp., *M.thermoautotrophicum, M.jannaschii, Pyrococcus furiosus, Pyrococcus horikoshii*
Bacteria	*A.aeolicus, C.trachomatis, Synechocystis* sp., *P.gingivalis, M.leprae, M.tuberculosis, B.subtilis, C.acetobutylicum, E.faecalis, M.genitalium, M.pneumoniae, S.pneumoniae, S.pyogenes, Rhizobium* sp., *R.capsulatus, S.aromaticivorans, N.gonorrhoeae, N.meningitidis, C.jejuni, H.pylori, E.coli, Y.pestis, H.influenzae, P.aeruginosa, B.burgdorferi, T.pallidum, D.radiodurans*
Additional Genomes on the public server at Integrated Genomics Inc.	*A.pernix, M.bovis, C.tepidum, S.typhi, T.maritima, A.actinomycetemcomitans, E.nidulans, Oryza sativa, A thaliana, R.prowazekii, P.abysii, C.pneumoniae, C.reinhardtii*

functional sub-systems, as a model for further refinement. For example, it is now possible to identify inconsistencies, potentially missing enzymes/ORFs, thereby refining the model. When a basic model has been created, a curator finally evaluates this model against biochemical data and phenotypes known from the literature. The models come in both textual and graphical representations, fully linked with all underlying data. We call this whole process metabolic reconstruction, and the main role of the WIT system is to support this effort.

To examine or curate a functional model of an organism, one can use functions such as: Compare assignments, Summary of asserted functions and pathways, Examine trimmed ortholog clusters, Examine COG/trimmed ortholog cluster relationships, Search for pathways by regular expression, Search ORF functions by regular expression, Search ORF sequences by similarity search, Find NCBI's MEDLINE-references by EC-number, Search EMP by EC-number, and Find common proteins for organisms. Chromosomal clustering of functionally related genes (14) is another powerful component of the system, which recently allowed us to propose a number of candidate ORFs for 'orphan' metabolic functions. Continuous integration of newly sequenced genomes increases the depth of functional description by a reiterative process.

GAPPED GENOMES IN WIT

An important feature of the WIT system is its emphasis on incomplete or gapped genomes. Algorithms used for gene assignments depend on the size of a dataset used to cluster properties of ORFs, whether it is chromosomal position or ortholog clustering based on bi-directional best hits. By the incorporation of gapped genomes, even the public version of WIT has integrated twice as much data as can be collected from only the completed genomes.

We believe that integrating systems like WIT can offer a solution for the problem of efficient use of incomplete sequence data. The gapped sequence contains a piece of almost every ORF, which allows the assignment of functions to almost all ORFs and the accurate reconstruction of the metabolism of the organism; good informatics can compensate for poorer sequence quality. A comparison of the results of analysis of the gapped genome of *Pseudomonas aeruginosa* with the complete genomes of *Escherichia coli* and *Bacillus subtilis* proves this statement (Table 2).

CONCLUSIONS

WIT has been designed to extract functional content from genome sequences and organize it into a coherent system, in order to facilitate post-sequencing experimental biology. The WIT system provides a set of local tools, which can be used to investigate functions of individual ORFs, based on similarities, motifs and various types of ORF clustering. It also generates overviews of functional subsystems and means to connect them into a complete picture of cellular functionality.

The WIT system is undergoing constant improvements, which can be traced in the PUMA–WIT–WIT2 line of development, and we believe that numerous further additions are needed to provide an adequate toolbox for the biological research community. Major directions of the ongoing WIT development are the following: (i) integration of structural data, which are currently underutilized in WIT; (ii) further development of the collection of functional maps and construction of more abstract scalable overviews, which should eventually cover all cellular functionality, and; (iii) development of a framework, which

Table 2. Comparison of the gapped *P.aeruginosa* genome with those of *E.coli* K-12 and *B.subtilis* 168

	Pseudomonas aeruginosa	*Escherichia coli*	*Bacillus subtilis*
Genome Size (Mb)	6.2	4.7	4.1
DNA assembled (%)	99	100	100
Total ORFs	5627	4289	4083
Assigned ORFs	4191	3499	3016
Asserted pathways	581	906	782
Missing assignments	133	102	178
No sequences	115	233	173

will integrate a flood of the differential display expression array data into the metabolic context.

ACKNOWLEDGEMENTS

Development of WIT was supported in part by the Office of Biological and Environmental Research, US Department of Energy, under Contract W-31-109-Eng-38, and in part by Integrated Genomics Inc.

REFERENCES

1. Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R., Bult,C.J., Tomb,J.F., Dougherty,B.A. and Merrick,J.M. (1995) *Science*, **269**, 496–512.
2. Kyrpides,N.C. (1999) *Bioinformatics*, **15**, 773–774.
3. Collins,F.S., Patrinos,A., Jordan,E., Chakravarti,A., Gesteland,R. and Walters,L. (1998) *Science*, **282**, 682–689.
4. Venter,J.C., Adams,M.D., Sutton,G.G., Kerlavage,A.R., Smith,H.O. and Hunkapiller,M. (1998) *Science*, **280**, 1540–1542.
5. Andrade,M.A., Brown,N.P., Leroy,C., Hoersch,S., de Daruvar,A., Reich,C., Franchini,A., Tamames,J., Valencia,A., Ouzounis,C. and Sander,C. (1999) *Bioinformatics*, **15**, 391–412.
6. Ogata,H., Goto,S., Sato,K., Fujibuchi,W., Bono,H. and Kanehisa,M. (1999) *Nucleic Acids Res.*, **27**, 29–34.
7. Mewes,H.W., Heumann,K., Kaps,A., Mayer,K., Pfeiffer,F., Stocker,S. and Frishman,D. (1999) *Nucleic Acids Res.*, **27**, 44–48. Updated article in this issue: *Nucleic Acids Res.* (2000), **28**, 37–40.
8. Benson,D.A., Boguski,M.S., Lipman,D.J., Ostell,J., Ouellette,B.F., Rapp,B.A. and Wheeler,D.L. (1999) *Nucleic Acids Res.*, **27**, 12–7. Updated article in this issue: *Nucleic Acids Res.* (2000), **28**, 15–18.
9. Bult,C.J., White,O., Olsen,G.J., Zhou,L., Fleischmann,R.D., Sutton,G.G., Blake,J.A., FitzGerald,L.M., Clayton,R.A., Gocayne,J.D., Kerlavage,A.R., Dougherty,B.A., Tomb,J.F., Adams,M.D., Reich,C.I., Overbeek,R., Kirkness,E.F., Weinstock,K.G., Merrick,J.M., Glodek,A., Scott,J.L., Geoghagen,N. and Venter,J.C. (1996) *Science*, **273**, 1058–1073.
10. Kaneko,T., Sato,S., Kotani,H., Tanaka,A., Asamizu,E., Nakamura,Y., Miyajima,N., Hirosawa,M., Sugiura,M., Sasamoto,S., Kimura,T., Hosouchi,T., Matsuno,A., Muraki,A., Nakazaki,N., Naruo,K., Okumura,S., Shimpo,S., Takeuchi,C., Wada,T., Watanabe,A., Yamada,M., Yasuda,M. and Tabata,S. (1996) *DNA Res.*, **3**, 185–209.
11. Vlcek,C., Paces,V., Maltsev,N., Haselkorn,R. and Fonstein,M. (1997) *Proc. Natl Acad. Sci. USA*, **94**, 9384–9388.
12. Selkov,E., Maltsev,N., Olsen,G.J., Overbeek,R. and Whitman,W.B. (1997) *Gene*, **197**, GC11–GC26.
13. Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) *Science*, **278**, 631–637.
14. Overbeek,R., Fonstein,M., D'Souza,M., Pusch,G.D. and Maltsev,N. (1999) *Proc. Natl Acad. Sci, USA*, **96**, 2896–2901.
15. Selkov,E., Basmanova,S., Gaasterland,T., Goryanin,I., Gretchkin,Y., Maltsev,N., Nenashev,V., Overbeek,R., Panyushkina,E., Pronevitch,L., Selkov,E.,Jr and Yunus,I. (1996) *Nucleic Acids Res.*, **24**, 26–28.
16. Selkov,E., Overbeek,R., Kogan,Y., Chu,L., Vonstein,V., Holmes,D., Silver,S., Haselkorn,R. and Fonstein,M. (1999) *Proc. Natl. Acad. Sci. USA.*, in press.
17. Badger,J.H. and Olsen,G.J. (1999) *Mol. Biol. Evol.*, **16**, 512–524.

Proc. Natl. Acad. Sci. USA
Vol. 85, pp. 2444-2448, April 1988
Biochemistry

APPENDIX B

Improved tools for biological sequence comparison

(amino acid/nucleic acid/data base searches/local similarity)

WILLIAM R. PEARSON* AND DAVID J. LIPMAN†

*Department of Biochemistry, University of Virginia, Charlottesville, VA 22908; and †Mathematical Research Branch, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892

Communicated by Gerald M. Rubin, December 2, 1987 (received for review September 17, 1987)

ABSTRACT We have developed three computer programs for comparisons of protein and DNA sequences. They can be used to search sequence data bases, evaluate similarity scores, and identify periodic structures based on local sequence similarity. The FASTA program is a more sensitive derivative of the FASTP program, which can be used to search protein or DNA sequence data bases and can compare a protein sequence to a DNA sequence data base by translating the DNA data base as it is searched. FASTA includes an additional step in the calculation of the initial pairwise similarity score that allows multiple regions of similarity to be joined to increase the score of related sequences. The RDF2 program can be used to evaluate the significance of similarity scores using a shuffling method that preserves local sequence composition. The LFASTA program can display all the regions of local similarity between two sequences with scores greater than a threshold, using the same scoring parameters and a similar alignment algorithm; these local similarities can be displayed as a "graphic matrix" plot or as individual alignments. In addition, these programs have been generalized to allow comparison of DNA or protein sequences based on a variety of alternative scoring matrices.

We have been developing tools for the analysis of protein and DNA sequence similarity that achieve a balance of sensitivity and selectivity on the one hand and speed and memory requirements on the other. Three years ago, we described the FASTP program for searching amino acid sequence data bases (1), which uses a rapid technique for finding identities shared between two sequences and exploits the biological constraints on molecular evolution. FASTP has decreased the time required to search the National Biomedical Research Foundation (NBRF) protein sequence data base by more than two orders of magnitude and has been used by many investigators to find biologically significant similarities to newly sequenced proteins. There is a trade-off between sensitivity and selectivity in biological sequence comparison: methods that can detect more distantly related sequences (increased sensitivity) frequently increase the similarity scores of unrelated sequences (decreased selectivity). In this paper we describe a new version of FASTP, FASTA, which uses an improved algorithm that increases sensitivity with a small loss of selectivity and a negligible decrease in speed. We have also developed a related program, LFASTA, for local similarity analyses of DNA or amino acid sequences. These programs run on commonly available microcomputers as well as on larger machines.

METHODS

The search algorithm we have developed proceeds through four steps in determining a score for pair-wise similarity. FASTP and FASTA achieve much of their speed and selectivity in the first step, by using a lookup table to locate all identities or groups of identities between two DNA or amino acid sequences during the first step of the comparison (2). The *ktup* parameter determines how many consecutive identities are required in a match. For example, if *ktup* = 4 for a DNA sequence comparison, only those identities that occur in a run of four consecutive matches are examined. In the first step, the 10 best diagonal regions are found using a simple formula based on the number of *ktup* matches and the distance between the matches without considering shorter runs of identities, conservative replacements, insertions, or deletions (1, 3).

In the second step of the comparison, we rescore these 10 regions using a scoring matrix that allows conservative replacements and runs of identities shorter than *ktup* to contribute to the similarity score. For protein sequences, this score is usually calculated using the PAM250 matrix (4), although scoring matrices based on the minimum number of base changes required for a replacement or on an alternative measure of similarity can also be used with FASTA. For each of these best diagonal regions, a subregion with maximal score is identified. We will refer to this region as the "initial region"; the best initial regions from Fig. 1*A* are shown in Fig. 1*B*.

The FASTP program uses the single best scoring initial region to characterize pair-wise similarity; the initial scores are used to rank the library sequences. FASTA goes one step further during a library search; it checks to see whether several initial regions may be joined together. Given the locations of the initial regions, their respective scores, and a "joining" penalty (analogous to a gap penalty), FASTA calculates an optimal alignment of initial regions as a combination of compatible regions with maximal score. FASTA uses the resulting score to rank the library sequences. We limit the degradation of selectivity by including in the optimization step only those initial regions whose scores are above a threshold. This process can be seen by comparing Fig. 1*B* with Fig. 1*C*. Fig. 1*B* shows the 10 highest scoring initial regions after rescoring with the PAM250 matrix; the best initial region reported by FASTP is marked with an asterisk. Fig. 1*C* shows an optimal subset of initial regions that can be joined to form a single alignment.

In the fourth step of the comparison, the highest scoring library sequences are aligned using a modification of the optimization method described by Needleman and Wunsch (5) and Smith and Waterman (6). This final comparison considers all possible alignments of the query and library sequence that fall within a band centered around the highest scoring initial region (Fig. 1*D*). With the FASTP program, optimization frequently improved the similarity scores of related sequences by factors of 2 or 3. Because FASTA calculates an initial similarity score based on an optimization of initial regions during the library search, the initial score is

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: NBRF, National Biomedical Research Foundation.

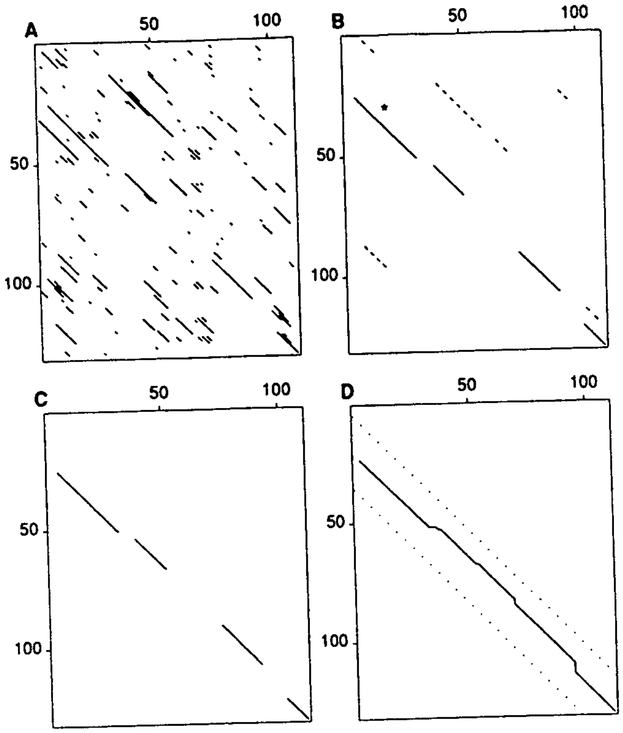


FIG. 1. Identification of sequence similarities by FASTA. The four steps used by the FASTA program to calculate the initial and optimal similarity scores between two sequences are shown. (*A*) Identify regions of identity. (*B*) Scan the regions using a scoring matrix and save the best initial regions. Initial regions with scores less than the joining threshold (27) are dashed. The asterisk denotes the highest scoring region reported by FASTP. (*C*) Optimally join initial regions with scores greater than a threshold. The solid lines denote regions that are joined to make up the optimized initial score. (*D*) Recalculate an optimized alignment centered around the highest scoring initial region. The dotted lines denote the bounds of the optimized alignment. The result of this alignment is reported as the optimized score.

much closer to the optimized score for many sequences. In fact, unlike FASTP, the FASTA method may yield initial scores that are higher than the corresponding optimized scores.

Local Similarity Analyses. Molecular biologists are often interested in the detection of similar subsequences within longer sequences. In contrast to FASTP and FASTA, which report only the one highest scoring alignment between two sequences, local sequence comparison tools can identify multiple alignments between smaller portions of two sequences. Local similarity searches can clearly show the results of gene duplications (see Fig. 2) or repeated structural features (see Fig. 3) and are frequently displayed using a "graphic matrix" plot (7), which allows one to detect regions of local similarity by eye. Optimal algorithms for sensitive local sequence comparison (6, 8, 9) can have tremendous computational requirements in time and memory, which make them impractical on microcomputers and, when comparing longer sequences, on larger machines as well.

The program for detecting local similarities, LFASTA, uses the same first two steps for finding initial regions that FASTA uses. However, instead of saving 10 initial regions, LFASTA saves all diagonal regions with similarity scores greater than a threshold. LFASTA and FASTA also differ in the construction of optimized alignments. Instead of focusing on a single region, LFASTA computes a local alignment for each initial region. Thus LFASTA considers all of the initial regions shown in Fig. 1*B*, instead of just the diagonal shown in Fig. 1*D*. Furthermore, LFASTA considers not only the band around each initial region but also potential sequence alignments for some distance before and after the initial region. Starting at the end of the initial region, an optimization (6) proceeds in the reverse direction until all possible alignment scores have gone to zero. The location of the maximal local similarity score in the reverse direction is then used to start a second optimization that proceeds in the forward direction. An optimal path starting from the forward maximum is then displayed (5). The local homologies can be displayed as sequence alignments (see Fig. 2*B*) or on a two-dimensional graphic matrix style plot (see Figs. 2*A* and 3).

Statistical Significance. The rapid sequence comparison algorithms we have developed also provide additional tools for evaluating the statistical significance of an alignment. There are approximately 5000 protein sequences, with 1.1 million amino acid residues, in the NBRF protein sequence library, and any computer program that searches the library by calculating a similarity score for each sequence in the library will find a highest scoring sequence, regardless of whether the alignment between the query and library sequence is biologically meaningful or not. Accompanying the previous version of FASTP was a program for the evaluation of statistical significance, RDF, which compares one sequence with randomly permuted versions of the potentially related sequence.

We have written a new version of RDF (RDF2) that has several improvements. (*i*) RDF2 calculates three scores for each shuffled sequence: one from the best single initial region (as found by FASTP), a second from the joined initial regions (used by FASTA), and a third from the optimized diagonal. (*ii*) RDF2 can be used to evaluate amino acid or DNA sequences and allows the user to specify the scoring matrix to be employed. Thus sequences found using the PAM250 scoring matrix can be evaluated using the identity or genetic code matrix. (*iii*) The user may specify either a global or local shuffle routine.

Locally biased amino acid or nucleotide composition is perhaps the most common reason for high similarity scores of dubious biological significance (10). High scoring alignments between query and library sequences may be due to patches of hydrophobic or charged amino acid residues or to A+T- or G+C-rich regions in DNA. A simple Monte Carlo shuffle analysis that constructs random sequences by taking each residue in one sequence and placing it randomly along the length of the new sequence will break up these patches of biased composition. As a result, the scores of the shuffled sequences may be much lower than those of the unshuffled sequence, and the sequences will appear to be related. Alternatively, shuffled sequences can be constructed by permuting small blocks of 10 or 20 residues so that, while the order of the sequence is destroyed, the local composition is not. By shuffling the residues within short blocks along the sequence, patches of G+C- or A+T-rich regions in DNA, for example, are undisturbed. Evaluating significance with a local shuffle is more stringent than the global approach, and there may be some circumstances in which both should be used in conjunction. Whereas two proteins that share a common evolutionary ancestor may have clearly significant similarity scores using either shuffling strategy, proteins related because of secondary structure or hydropathic profile may have similarity scores whose significance decreases dramatically when the results of global and local shuffling are compared.

Implementation. The FASTA/LFASTA package of sequence analysis tools is written in the C programming language and has been implemented under the Unix, VAX/VMS, and IBM PC DOS operating systems. Versions of the program that run on the IBM PC are limited to query se-

Table 1. FASTA and FASTP initial scores of the T-cell receptor (RWMSAV) versus the NBRF data base

		Initial score	
NBRF code	Sequence	FASTA	FASTP
RWHUAV	T-cell receptor α chain	155	98
K1HURE	Ig κ chain V-I region	127	111
KVMS50	Ig κ chain V region	149	62
KVMSM6	Ig κ chain precursor V regions	141	64
KVRB29	Ig κ chain V region	126	54
L3HUSH	Ig λ chain V-III region	90	47
KVMS41	Ig κ chain precursor V region	87	87
RWMSBV	T-cell receptor β-chain precursor	94	94
RWHUVY	T-cell receptor β-chain precursor	91	59
RWHUGV	T-cell receptor γ-chain precursor	87	61
RWHUT4	T-cell surface glycoprotein T4	86	63
RWMSVB	T-cell receptor γ-chain precursor	71	41
HVMS44	Ig heavy-chain V region	67	36
G1HUDW	Ig heavy-chain V-II region	62	35

The average FASTP score = 26.1 ± 6.8 (mean ± SD). The average FASTA score = 26.2 ± 7.2 (mean ± SD). The mean and SD were computed excluding scores >54. V, Variable.

quences of 2000 residues; library sequences can be any length. Copies of the program are available from the authors.

Although FASTA and LFASTA were designed for protein and DNA sequence comparison, they use a general method that can be applied to any alphabet with arbitrary match/mismatch scoring values. All the scoring parameters, including match/mismatch values, values for the first residue in a gap and subsequent residues in the gap, and other parameters that control the number of sequences to be saved and the histogram intervals, can be specified without changing the program.

EXAMPLES

Comparison of FASTA with FASTP. To demonstrate the superiority of the FASTA method for computing the initial score, we compared the protein sequence of a T-cell receptor α chain (NBRF code RWMSAV) with all sequences in the NBRF protein data base[‡] and computed initial scores with both the present and previous methods. The T-cell receptor is a member of the immunoglobulin superfamily; in Release 12.0 of the data base, this superfamily has 203 members. FASTP placed 160 immunoglobulin superfamily sequences in the 200 top-scoring sequences; 57 related sequences received initial scores less than four standard deviations above the mean score. FASTA placed 180 superfamily members in the 200 top-scoring sequences; only 20 related sequences scored below four standard deviations above the mean. Table 1 contains specific examples from this data base search. Although there is often little difference in the two methods, this example shows that in a number of cases the new method obtains significantly higher scores between related sequences.

Nucleic Acid Data Base Search. FASTA can also be used to search DNA sequence data bases, either by comparing a DNA query sequence to the DNA library or by comparing an amino acid query sequence to the DNA library by translating each library DNA sequence in all six possible reading frames. We compared the 660-nucleotide rat transforming growth factor type α mRNA (GenBank locus RATTGFA) with all the mammalian sequences in Release 48 of GenBank[§]. We set *ktup* = 4 (see *Methods*), and the search was completed in under 15 min on an IBM PCAT microcomputer. The 10 top-scoring library sequences are shown in Table 2. Although it can be seen that the 3 top-scoring sequences are clearly related to RATTGFA, there are other high-scoring sequences that are probably not related, and the mouse epidermal growth factor, found in the translated data base search (Table 3), is not found among the top-scoring sequences.

Table 2. DNA data base search of rat transforming growth factor (RATTGFA) versus mammalian sequences

GenBank		Score	
locus	Sequence	Initial	Optimized
HUMTFGAM	Human TGF mRNA	1336	1618
HUMTGFA2	Human TGF gene (exon 2)	354	366
HUMTGFA1	Human TGF gene (5′ end)	224	381
MUSRGEB3	Mouse 18S–5.8S–28S rRNA gene	140	107
MUSRGE52	Mouse 18S–5.8S–28S rRNA gene	140	107
MUSMHDD	MHC class I H-2D	122	78
HUMMETIF1	Metallothionein (MT)I$_F$ gene	116	92
MUSRGLP	45S rRNA (5′ end)	115	83
HUMPS2	pS2 mRNA	105	106
MUSC1A11	α-1 type I procollagen	86	89

The 10 sequences having the highest initial scores are given. TGF, transforming growth factor; MHC, major histocompatibility complex.

To further examine the similarity detected between RATTGFA and MUSRGEB3, a mouse rRNA gene cluster, we used the RDF2 program for Monte Carlo analysis of statistical significance (the window for local shuffling was set to 10 bases). Of the 50 shuffled comparisons (data not shown), 1 obtained an initial score greater than 140 (the observed initial score), and 9 shuffled sequences obtained optimized scores greater than 107 (the observed optimized score). Therefore, the similarity between RATTGFA and MUSRGEB3 is unlikely to be significant.

Translated Nucleic Acid Data Base Search. When searching for sequences that encode proteins, amino acid sequence comparisons are substantially more sensitive than DNA sequence comparisons because one can use scoring matrices like the PAM250 matrix that discriminate between conservative and nonconservative substitutions. A variant of FASTA, TFASTA, can be used to compare a protein sequence to a DNA sequence library; it translates the DNA sequences into each of six possible reading frames "on-the-fly." TFASTA translates the DNA sequences from beginning to end; it includes both intron and exon sequences in the translated protein sequence; termination codons are translated into unknown (X) amino acids. Table 3 shows the results of a translating search of the mammalian sequences in the GenBank DNA data base using the RATTGFA protein sequence as the query and *ktup* = 1. In the translated search, the mouse epidermal growth factor now obtains an initial score higher than any unrelated sequences; however, HUMTGFA1, which was found in the DNA data base search but only contains 13 translated codons, is no longer among the top scoring sequences.

Local Similarities. Fig. 2 displays the output of a local similarity analysis (*ktup* = 4) of CHPHBA1M, a chimpanzee α1-globin mRNA, and RABHBAPT, a rabbit α-globin gene, including the complete coding sequence and a flanking pseudo-θ_1-globin gene. LFASTA can either display a graphic matrix style plot of the local homologies (Fig. 2*A*) or the alignments themselves (Fig. 2*B*). The right-most three alignments (Fig. 2*A*) match the corresponding regions of the mRNA to exon subsequences from the pseudogene. We note that the FASTA initial score for the comparison of CHPH-

[‡] Protein Identification Resource (1987) Protein Sequence Database (Natl. Biomed. Res. Found., Washington, DC), Release 12.

[§] EMBL/GenBank Genetic Sequence Database (1987) (Intelligenetics, Mountain View, CA), Tape Release 48.

Table 3. Translated DNA data base search of rat transforming growth factor (RATTGFA) versus mammalian sequences

GenBank locus	Sequence	Frame	Score	
			Initial	Optimized
RATTGFA	Rat TGF type *α*	1	816	816
HUMTGFAM	Human TGF mRNA	2	671	770
HUMTGFA2	Human TGF gene	1	204	205
MUSEGF	Mouse EGF mRNA	3	93	129
MUSMHAB3	Mouse MHC class II H2-IA_{β}	1	91	58
MUSIGCD17	Mouse Ig germ-line DJC region	3′	85	48
HUMESTR	Human estrogen receptor	3	83	65
RATINS1	Rat insulin 1 (*Ins-1*) gene	2	81	63
MUSTHYS1	Mouse thymidylate synthase	2	80	63
HUMPNU3	Human purine nucleoside phosphorylase	1′	80	52

The 10 sequences having the highest initial scores are given. TGF, transforming growth factor; EGF, epidermal growth factor; D, diversity; J, joining; C, constant; MHC, major histocompatibility complex.

BA1M and RABHBAPT would be based on the three globin gene exons, while the FASTP initial score would be based on a single conserved exon.

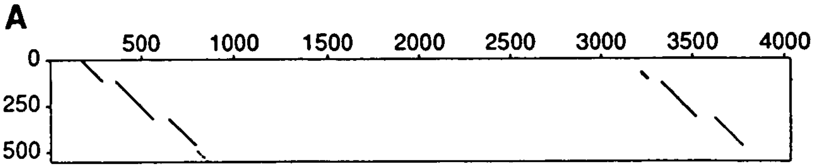
The Smith–Waterman optimization used in the LFASTA program allows the detection of more subtle features than can be detected by the eye using a graphic matrix plot, because the path traced is locally optimal, even though it may only have a slightly higher density of identities and conservative replacements. Fig. 3 shows a plot from a local similarity self-comparison of the myosin heavy chain from the nematode *Caenorhabditis elegans* (MWKW) using the PAM250 matrix. The amino-terminal half of the molecule forms a large globular head without any periodic structure; the solid line down the main diagonal represents the expected identity of the sequence with itself. The symmetrical parallel lines along the carboxyl-terminal half of the molecule correspond to the 28-residue repeat responsible for the *α*-helical coiled-coil structure of the rod segment.

DISCUSSION

In searching a data base, one is attempting to measure relatedness; in aligning two homologous sequences, one is trying to choose the most likely set of mutations since their divergence from a common ancestral sequence. Thus any tool for the analysis of sequence similarities must contain within it an implicit model of molecular evolution. An algorithm that guarantees the optimality of its alignments based on a set of scoring rules must be judged on how well these rules fit our current understanding of the process of molecular evolution. Algorithms that sacrifice realism to achieve greater efficiency, regardless of their mathematical rigor, require careful empirical evaluation.

Even though the tools we have developed use rigorous algorithms at each step and incorporate a realistic model of evolution, their hierarchical nature make them heuristic. The original FASTP program has had the benefit of extensive use and evaluation by a wide variety of scientists. The FASTA program exploits refinements of the previous approach that result in a significant improvement in sensitivity. The LFASTA local similarity analysis program is also a logical extension of the FASTP approach.

Because of the trade-offs between sensitivity and selectivity in data base searches, the results of any search, and particularly those that result in alignment scores that are not clearly separated from the distribution of all library sequence



B

```
               10        20        30        40        50        60
CHPHBA GACTCAGAAAGAACCCACCATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAAGGCCG
       :::: :::: ::: X:::::::::::::::::: :: :::::::::::: ::::: : :
RABHBA GACTGAGAAGGAA-CCACCATGGTGCTGTCTCCCGCTGACAAGACCAACATCAAGACTG
              180        190       200       210       220

               70        80        90       100       110
CHPHBA CCTGGGGTAAGGTCGGCGCGCACGCTGGCGAGTATGGTGCGGAGGCCCTGGAGAGG
       ::::::  ::: :::::   :::: :::::::::::: :: :::::: :::::::X
RABHBA CCTGGGAAAAGATCGGCAGCCACGGTGGCGAGTATGGCGCCGAGGCCGTGGAGAGG
      230       240       250       260       270       280
```

FIG. 2. Local comparison of an *α*-globin mRNA sequence with an *α*-globin gene cluster. An ape α_1-globin mRNA sequence (GenBank sequence CHPHBA1M) was compared with a rabbit *α*-globin gene sequence (RABHBAPT) containing a second pseudo-*θ*-globin gene using the LFASTA program. (*A*) A plot of the homologous regions shared by the two sequences. (*B*) One of the alignments between the mRNA sequence and the rabbit *α*-globin gene (nucleotides 171–855). Three other alignments between the mRNA sequence and the *α*-globin gene and three alignments between the pseudo-*θ*-globin gene (nucleotides 3200–3770) were calculated but are not shown. There is 84.3% identity in the 115 nucleotide overlap. The initial region and optimized scores using LFASTA are 284 and 304, respectively. X denotes the ends of the initial region found by LFASTA.

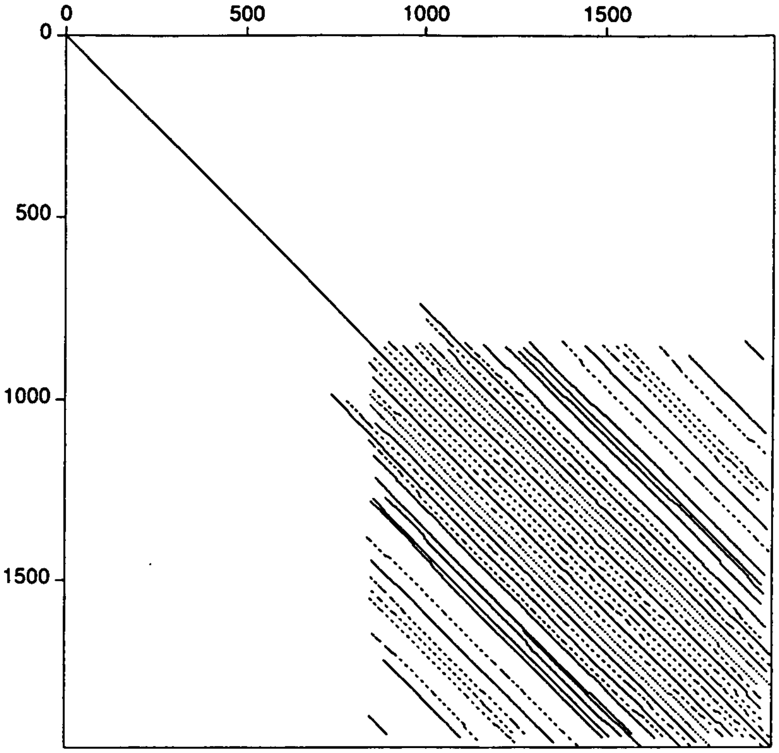


FIG. 3. Repeated structure in the myosin heavy chain. LFASTA was used to compare the *Caenorhabditis elegans* myosin heavy chain protein sequence (NBRF code MWKW) with itself using the PAM250 scoring matrix. The solid, dashed, and dotted lines denote decreasing similarity scores. The solid lines had initial region scores greater than 80 and optimized local scores greater than 150; the longer dashed lines had initial region and optimized local scores greater than 65 and 120, respectively, and the shorter dashed lines had initial region and optimized local scores greater than 50 and 100, respectively. Homologous regions with lower scores are plotted with dots.

scores, must be carefully evaluated (1, 11). The Monte Carlo analysis of statistical significance provided by a program such as RDF2 can often be critical in evaluating a borderline similarity. Previously we suggested ranges of z values [(observed score − mean of shuffled scores)/standard deviation of shuffled scores] corresponding to approximate significance levels. However the z values determined in a Monte Carlo analysis become less useful as the distribution of shuffled scores diverges from a normal distribution, as is found with FASTA. Therefore, we now focus on the highest scores of the shuffled sequences. For example, if in 50 shuffled comparisons, several random scores are as high or higher than the observed score, then the observed similarity is not a particularly unlikely event. One can have more confidence if in 200 shuffled comparisons, no random score approaches the observed score. In general, our experience has led us to be conservative in evaluating an observed similarity in an unlikely biological context.

These programs provide a group of sequence analysis tools that use a consistent measure for scoring similarity and constructing alignments. FASTA, RDF2, and LFASTA all use the same scoring matrices and similar alignment algorithms, so that potentially related library sequences discovered after the search of a sequence data base can be evaluated further from a variety of perspectives. In addition, LFASTA can also show alternative alignments between sequences with periodic structures or duplications.

1. Lipman, D. J. & Pearson, W. R. (1985) *Science* **227**, 1435–1441.
2. Dumas, J. P. & Ninio, J. (1982) *Nucleic Acids Res.* **10**, 197–206.
3. Wilbur, W. J. & Lipman, D. J. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 726–730.
4. Dayhoff, M., Schwartz, R. M. & Orcutt, B. C. (1978) in *Atlas of Protein Sequence and Structure*, ed. Dayhoff, M. (Natl. Biomed. Res. Found., Silver Spring, MD), Vol. 5, Suppl. 3, pp. 345–352.
5. Needleman, S. & Wunsch, C. (1970) *J. Mol. Biol.* **48**, 444–453.
6. Smith, T. & Waterman, M. S. (1981) *J. Mol. Biol.* **147**, 195–197.
7. Maizel, J. & Lenk, R. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 7665–7669.
8. Goad, W. & Kanehisa, M. (1982) *Nucleic Acids Res.* **10**, 247–263.
9. Sellers, P. H. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 3041.
10. Lipman, D. J., Wilbur, W. J., Smith, T. F. & Waterman, M. S. (1984) *Nucleic Acids Res.* **12**, 215–226.
11. Doolittle, R. (1981) *Science* **214**, 149–159.

Appendix C

Beyond complete genomes: from sequence to structure and function

Eugene V Koonin*, Roman L Tatusov and Michael Y Galperin

Computer analysis of complete prokaryotic genomes shows that microbial proteins are in general highly conserved – ~70% of them contain ancient conserved regions. This allows us to delineate families of orthologs across a wide phylogenetic range and, in many cases, predict protein functions with considerable precision. Sequence database searches using newly developed, sensitive algorithms result in the unification of such orthologous families into larger superfamilies sharing common sequence motifs. For many of these superfamilies, prediction of the structural fold and specific amino acid residues involved in enzymatic catalysis is possible. Taken together, sequence and structure comparisons provide a powerful methodology that can successfully complement traditional experimental approaches.

Addresses
National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA
*e-mail: koonin@ncbi.nlm.nih.gov
Correspondence: Eugene V Koonin

Current Opinion in Structural Biology 1998, 8:355–363

http://biomednet.com/elecref/0959440X00800355

Abbreviations

COGs	clusters of orthologous groups
HAD	haloacid dehalogenase

Introduction

The determination of the complete genome sequences of several bacteria and archea and one eukaryote [1–6,7••–12••] marked the beginning of a new age in biology. For the first time, we can take a look at the complete set of proteins present in the cells of each particular organism and try to identify the proteins responsible for each cellular function. In cases where no known proteins can be found to perform a particular task, the most likely substitutes can be predicted from the set of unassigned gene products. Clearly this can be done only by analysis of complete genomes, as partial sequences do not allow us to ascertain that certain proteins are not encoded in a given genome [13]. These new approaches are gradually changing our understanding of a variety of biological phenomena. As the number of sequenced genomes is expected to grow exponentially for the next few years, their impact on different biological disciplines will increase. We have recently discussed the implications of the complete genomes for microbial evolution [14]. Here we consider the effect of the genome revolution, together with the improving methods for sequence analysis, on our ability to predict and understand protein structure and function.

Towards a natural taxonomy of proteins and protein families

The numerous genome sequencing projects have resulted in a rapid growth of protein databases (see, e.g. [15]). In contrast to the pre-genome era, when researchers typically chose to clone and sequence genes with documented functional roles, we are now getting many protein sequences whose functions are not known. This presents a challenge to extract the most from these sequences in terms of salient features of the encoded proteins, for example to classify them according to their homologous relationships, and to predict their possible catalytic activities and/or cellular functions, three-dimensional (3D) structures and evolutionary origin.

Protein classifications, pioneered by Dayhoff and her coworkers, have historically been based on sequence alignments. Similar proteins formed families, which were combined into superfamilies [16]. This approach, continued in the PIR database [17], proved extremely popular. However, even PIR superfamilies often unite closely related proteins and more distant relationships are being missed. Other protein databases, such as PROSITE [18], PRINTS [19], Pfam [20], and ProDom [21], group proteins on the basis of conserved sequence motifs and, generally, contain much more diverse protein families. Structural comparisons of proteins, implemented in FSSP, CATH and SCOP databases, offer yet another approach to protein classification [22–24]. SCOP superfamilies, for example, unite proteins that have some similarities in their 3D structures, but often no detectable sequence similarity [25]. Thus, in the absence of clear sequence or structural similarities, the criteria for inclusion of distantly related proteins into a family (or superfamily) become increasingly arbitrary.

With the inception of extensive genome sequencing, it has become possible to classify genes and proteins on a different principle, namely by delineating families of paralogs — related genes within the same genome [26,27]. Such analyses have revealed a complex hierarchical organization of paralogous families in each of the studied genomes and produced at least two generalizations: first, the fraction of genes that belong to families of paralogs increases with the increase of the total number of genes in a genome: from ~25% in the minimal genome of *Mycoplasma genitalium* to >50% in the large (for a prokaryote) *Escherichia coli* genome; second, the largest superfamilies of paralogs are mostly the same in all genomes [28–33].

Knowledge of all the protein sequences from multiple complete genomes (Table 1) allows us to redefine the entire

Table 1

Protein families and 3D structures in complete genomes.

Species	Proteins encoded in the genome*		COGs found (% total)	3D structures	
	Total number	Belong to COGs† (% total)		In PDB	Predicted‡
Escherichia coli	4289	2003 (47%)	821 (95%)	240	867
Haemophilus influenzae	1717	979 (57%)	658 (77%)	2	267
Helicobacter pylori	1566	841 (54%)	617 (72%)	0	169
Synechocystis sp.	3169	1551 (49%)	796 (93%)	2	431
Borrelia burgdorferi	850	483 (57%)	363 (42%)	0	105
Bacillus subtilis	4100	1945 (47%)	732 (85%)	12	578
Mycoplasma genitalium	467	341 (75%)	290 (34%)	0	75/103
Mycoplasma pneumoniae	677	378 (56%)	309 (36%)	0	78
Methanococcus jannaschii	1715	830 (48%)	498 (58%)	0	170
Methanobacterium thermoautotrophicum	1869	897 (46%)	484 (56%)	0	199
Archaeoglobus fulgidus	2407	1131 (47%)	512 (60%)	0	290
Saccharomyces cerevisiae	5932	1736 (29%)	577,(67%)	45	846
Caenorhabditis elegans	12,178	2172 (18%)	466 (54%)	2	NA

*The numbers are from the latest updates in the GenBank genome division (ftp://ncbi.nlm.nih.gov/genbank/genomes). *C. elegans* genome is about 85% complete; the data are from Wormpep12 (www.sanger.ac.uk/Projects/C_elegans/wormpep). †Based on the set of 860 COGs, obtained by adding *H. pylori* proteins to the original set of 720 COGs [37••]. ‡The numbers are from the PEDANT database [53•], calculated by comparing the protein set encoded in each genome to the PDB using FASTA with cutoff score of 120; the second figure for *M. genitalium* is from [54•]; the data for *C. elegans* are not available.

problem of protein classification. Since the fraction of proteins conserved over large phylogenetic distances (ancient conserved domains) appears to be nearly constant at ~70% in all prokaryotic genomes [34•], it becomes feasible to replace more or less arbitrary clustering of proteins by similarity with consistent groups in which the evolutionary relationships between the members are specifically defined. Such a classification of proteins can provide a framework for evolutionary studies and for rapid, largely automatic, functional annotation of newly sequenced genomes.

Several classifications of homologous proteins encoded in complete genomes have been produced, based on all-against-all protein sequence comparisons [35,36,37••]. Each of these projects is aimed at the identification of orthologs, that is direct counterparts in different genomes, connected by an uninterrupted line of vertical descent and typically retaining their physiological function [26,27]. In particular, the system of clusters of orthologous groups (COGs) was designed to accommodate the vastly different evolution rates observed for different genes [37••]. The COGs construction procedure identifies the closest homologs in each of the sequenced genomes for each protein, even if the similarity is fairly low and not statistically significant by itself. The approach to the identification of COGs was built upon the transitivity of orthologous relationships, that is the simple notion that any group of at least three genes from distant genomes, which are more similar to each other than they are to any other genes from the same genomes, is most likely to belong to an orthologous family. Clearly, this is a probabilistic assumption based on a 'weak molecular clock concept', which posits that orthologs are more similar to each other than they are to paralogs with different, even if related, functions. This assumption, however, seems to hold true in cases where we have reasons to accept orthology on functional grounds (for example, aminoacyl-tRNA synthetases or ribosomal proteins). Orthology is not necessarily a one-to-one relationship, as in cases of lineage-specific duplications, orthology can only be established between families of paralogous genes. Such complex relationships require caution in the functional interpretation of the phylogenetic classification of proteins. Nevertheless, about 60% of the original set of 720 COGs [37••] are simple families, with no paralogs or with paralogs from one lineage only, suggesting the possibility of straightforward transfer of functional information from functionally characterized genes from model systems such as *E. coli* and yeast to those from poorly characterized genomes.

The utility of this system of protein classification was tested on several newly sequenced bacterial, archeal and eukaryotic genomes. Interestingly, with the only exception of the minimal genome of *M. genitalium*, the fraction of the proteins that belong to the COGs — ancient families conserved across a wide phylogenetic range — is about the same and very close to 50% for all prokaryotic genomes (Table 1). This is clearly compatible with the previous estimate that about 70% of the proteins encoded in each genome contain ancient conserved regions. The fraction of the proteins included in the COGs is at this time lower, which is evidently due to the requirement for three distant lineages to be included, and to the limited number of species in the first instalment of the COGs. There is little doubt that with new genomes added, the number of COGs will asymptotically approach the total number of ancient conserved regions. By contrast, this fraction is much lower for eukaryotic genomes, indicating the prevalence of eukaryote-specific families.

Comparison of the new protein sets with the COGs resulted in a number of functional predictions for previously uncharacterized proteins. Even for the *Helicobacter pylori* proteins, most of which show highly significant similarity to homologs from *E. coli* and other bacteria and have been described in considerable detail [8••], predictions were made in more than 100 cases (http://www.ncbi.nlm.nih/COG); function was also predicted for a number of archeal and worm proteins (EV Koonin, RL Tatusov, MY Galperin, unpublished data).

Missing gene families and evolution of metabolic pathways

Comparative analysis of the available complete genomes shows that metabolic diversity generally correlates with genome size. Parasitic bacteria import a variety of metabolites, which allows them to shed genes encoding enzymes for many or even most of the metabolic pathways [1–3, 8••,33,38]. In contrast, all cells have to rely on their own gene products for performing such essential functions as genome expression, replication and repair, and membrane biogenesis and others. These tasks alone require at least about 200 genes [13,37••].

Given complete genome sequences, classification of proteins into orthologous groups provides a convenient way to systematically survey the protein families present or absent in a genome and to identify the metabolic pathways that are likely to be operative in the organism analyzed. When some of the required enzymes cannot be found in the genome, the respective pathways are either not operative, or use other, unrelated, proteins to catalyze the missing steps (see [39]). An example of such an analysis, which included superposition of the phylogenetic patterns derived from the COGs [37••], over the scheme of glycolysis, reveals several interesting trends (Figure 1). Glycolysis includes three reactions that in different species are catalyzed by non-orthologous enzymes, namely phosphofructokinases, aldolases and phosphoglycerate mutases. Interestingly, the second phosphofructokinase in *E. coli*, encoded by the *pfkB* gene, has apparently been recruited from a ubiquitous family of ribokinase-like sugar kinases. The ribokinase COG seems to be an example of a complex family in which the exact orthologous connections are not always easy to trace. In particular, even though PfkB formally belongs to the COG, there seems to be no actual ortholog of it in other genomes. Thus *H. pylori* does not encode a phosphofructokinase at all, although it has genes for other kinases of the ribokinase family and, accordingly, is represented in the respective COG (Figure 1).

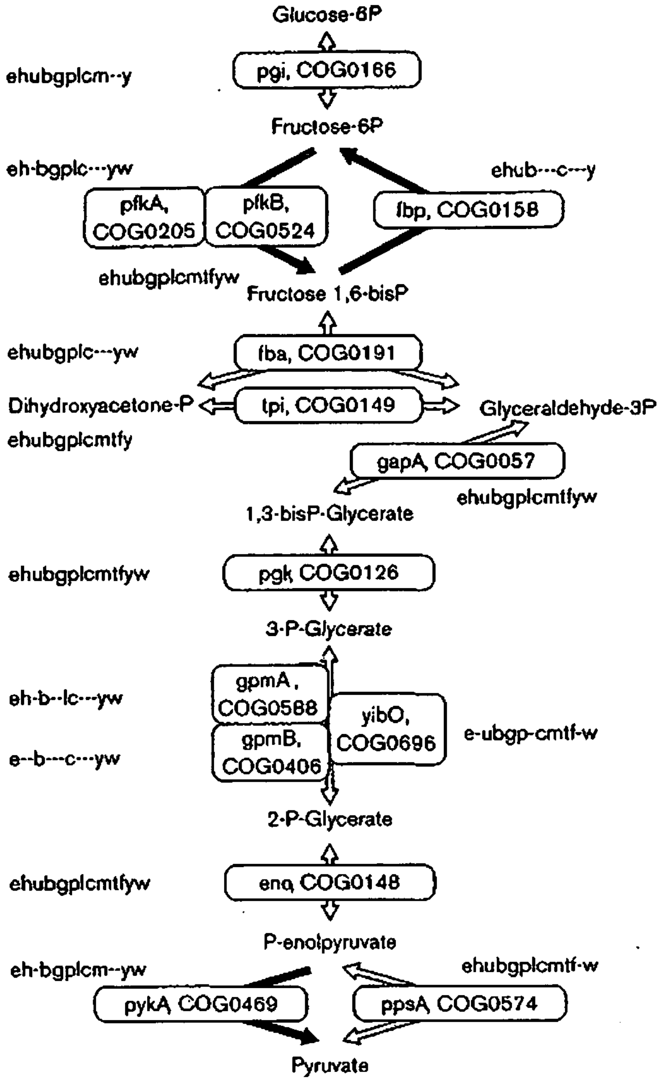
A remarkable case of non-orthologous gene displacement involves two unrelated forms of phosphoglycerate mutase, the 2,3-bisphosphoglycerate (BPG)-dependent and the BPG-independent one. While *H. influenzae* and *Borrelia burgdorferi* encode only the BPG-dependent form, and *H. pylori*, mycoplasmas, and archea encode only the BPG-independent form (see [40]), free-living bacteria such as *E. coli*, *Bacillus subtilis* and *Synechocystis* sp. possess genes coding for both these forms, with two paralogs of the BPG-dependent one (Figure 1). Phosphofructokinase, aldolase and fructose bisphosphatase genes are all missing in the archea (Figure 1), in accordance with the experimental data [41]. This is consistent with the idea that glycolysis originally evolved as a biosynthetic pathway, containing only the lower (tri-carbon) part [42].

Systematic identification of missing links in functional systems in organisms for which complete genome sequences are available is probably the most important application of protein family classification. Conspicuous gaps in the *H. pylori* metabolism became apparent from the COG analysis, suggesting major revisions to the general scheme of the central metabolic pathways in this bacterium (Table 2). In particular, unlike most other bacteria (and all with completely sequenced genomes), *H. pylori* seems to possess neither glycolysis nor the pentose phosphate shunt, the Entner-Doudoroff pathway being the only major route of sugar catabolism. Indeed, sugar fermentation, resulting in intracellular acid production, would be an additional burden on the pH maintenance mechanism in this bacterium, which has to survive in an external pH of 2–3. By contrast, gluconeogenesis, which converts organic acids into sugars required for nucleic acid and peptidoglycan biosynthesis and thus removes H^+ from the cytoplasm, appears to be fully functional in *H. pylori*. For the purpose of energy production, *H. pylori* apparently depends on amino acid fermentation, which causes alkalinization of the cytoplasm and thus relieves part of the problem of pH maintenance. Amino acids and oligopeptides that serve as substrates for this fermentation are produced by gastric proteolysis and transported by readily identifiable permeases.

From genomes and families to superfamilies and folds

Classification systems aimed at the identification of families of orthologs make no attempt to capture the more subtle conserved motifs in proteins, which reflect ancient relationships at the level of superfamilies and frequently are critically important for understanding protein functions and structures [43,44]. Computer methods for the detection of such motifs and delineation of superfamilies have lately progressed significantly through programs such as BLIMPS/MULTIMAT [45], Probe [46], and PSI-BLAST [47••], which combine pairwise sequence comparisons with profile analysis. PSI-BLAST, in particular, has proved to be a powerful tool for the detection of subtle sequence motifs, resulting in the discovery of a number of unsuspected superfamily relationships [47••,48•]. Furthermore, one of the perhaps under-appreciated benefits of the accumulation of genomic sequences is the greatly improved capacity to identify even very subtle sequence similarities due to

Figure 1



Current Opinion in Structural Biology

Glycolytic enzymes in organisms with completely sequenced genomes. The enzymes are listed under *E. coli* gene names. The COG numbers are as in COG database (www.ncbi.nlm.nih.gov/COG, [37••]) (where available). Shaded arrows indicate reversible reactions, black arrows practically irreversible ones. Phosphoenolpyruvate synthase-catalyzed reaction in the direction of phosphoenolpyruvate hydrolysis has been demonstrated *in vitro*. Phylogenetic patterns are: e, *Escherichia coli*; h, *Haemophilus influenzae*; u, *Helicobacter pylori*; b, *Bacillus subtilis*; g, *Mycoplasma genitalium*; p, *Mycoplasma pneumoniae*; l, *Borrelia burgdorferi*; c, *Synechocystis* sp.; m, *Methanococcus jannaschii*; t, *Methanobacterium thermoautotrophicum*; f, *Archaeoglobus fulgidus*; y, *Saccharomyces cerevisiae*; w, *Caenorhabditis elegans*.

the increasingly uniform population of the protein universe by these relatively unbiased sequence sets, of which the new methods for sequence analysis mentioned above can take advantage [49•].

In the past year, we have seen the identification or significant extension of a number of protein superfamilies; some examples, with the distribution among complete genomes, are shown in Table 3. Most of these superfamilies are universally found in all genomes, with the counts more or less proportional to the total number of genes in the genome. Some expansions are, however, remarkable, such as, for example, urease-related hydrolases and ATP-grasp domains in the archea, and HAD superfamily hydrolases in *E. coli* and *B. subtilis* (Table 3). In certain cases, the phylogenetic distribution of a superfamily immediately suggests major evolutionary events. Thus the BRCT domain is present in a single copy in the DNA ligase of all bacteria (with one additional copy found only in *Synechocystis*), is missing in the archea, and is dramatically expanded in its distribution in the eukaryotes (Table 3). The most obvious interpretation of this distribution is that this domain has entered the eukaryotic world by horizontal gene transfer from bacteria and has undergone exten-

Table 2

Genes and pathways missing in Helicobacter pylori.

Enzyme activity	*E. coli* gene	COG number	Status in *H. pylori*	Implications for *H. pylori* metabolism
Phosphofructokinase	*pfkA*	COG0206	Missing	Absence of the two key glycolytic enzymes shows that Embden-Meyerhof pathway is not functional in *H. pylori*. Gluconeogenesis enzymes, bypassing these reactions, fructose bisphosphatase (HP1385) and phosphoenolpyruvate synthase (HP0121), are present in *H. pylori*, allowing it to produce sugars required for peptidoglycan biosynthesis.
	pfkB	COG0525	Present (ribokinase)	
Pyruvate kinase	*pykA* *pykF*	COG0470	Missing	
6-phosphogluconate dehydrogenase	*gnd*	COG0360	Missing	Pentose phosphate pathway is also not functional. Even though *H. pylori* has a ribose 5-phosphate isomerase encoded by an ortholog of the E. coli *rpiB*, no gene coding for 6-phosphogluconate dehydrogenase could be identified. The only saccharolytic pathway in *H. pylori* appears to be the Entner-Doudoroff pathway.
Ribose 5-phosphate isomerase	*rpiA*	COG0120	Missing	
Lipoate synthase	*lipA*	COG0318	Missing	Pyruvate dehydrogenase complex is absent in *H. pylori*; acetate kinase and phosphotransacetylase are not functional. Pyruvate-ferredoxin oxidoreductase is the only acetyl-CoA-producing enzyme in *H. pylori*.
Lipoate-protein ligase	*lplA*	COG0411	Missing	
	lipB	COG0319	Missing	
Dihydrolipoamide acyltransferase	*aceF*	COG0510	Missing	
Acetate kinase	*ackA*	COG0280	Disrupted by a frameshift	
Phospho-transacetylase	*pta*	COG0278	Disrupted by frameshifts	
Enzymes of purine biosynthesis	*purF*	COG0034	Missing	*De novo* purine biosynthesis is absent in *H. pylori*, and it has to obtain purines from the host. HP1185 appears to be the best candidate for the purine permease, as it is the only *H. pylori* protein, similar to *E. coli* PurP.
	purD	COG0151	Inactivated by mutations	
	purN	COG0299	Missing	
	purT	COG0027	Missing	
	purL_1	COG0046	Missing	
	purL_2	COG0047	Missing	On the other hand, *H. pylori* encodes the enzymes for AMP and GMP synthesis from IMP and their interconversion. Therefore, it can survive on any of these purines.
	purM	COG0150	Missing	
	purK	COG0026	Missing	
	purE	COG0041	Missing	
	purC	COG0152	Missing	
	purH	COG0138	Missing	
	purA	COG0104	Present	
	purB	COG0015	Present	
	guaB	COG0516	Present	
	guaA_1	COG0518	Present	
	guaA_2	COG0519	Present	

sive duplication with divergence in the eukaryotes. The expansion of this domain into a number of eukaryotic proteins involved in cell-cycle control [50••,51] may have been critical for the very establishment of these systems.

With the current acceleration in protein structure determination [22,24], a superfamily identified by sequence comparison more and more frequently extends to include proteins with known 3D structure and/or well-characterized catalytic mechanism (Table 3). Such findings are sometimes most illuminating as they immediately result in the prediction of the structural fold, the structure of the active center, and possibly also the catalytic mechanism for a wide variety of diverse proteins comprising the superfamily. This is illustrated by the recent prediction of the structure and the catalytic amino acid residues for P-ATPases, which remained elusive in spite of a long history of studies, on the basis of the sequence motifs shared with haloacid dehalogenases [52•].

Assignment of the gene products to structural folds and families with maximal attainable precision is arguably one of the foremost tasks of genome analysis after the sequencing phase. The number of structures that have been determined experimentally is negligible for almost all genomes, with the exception of *E. coli* (where it is still rather a small fraction) (Table 1). A database search with a deliberately conservative similarity cut-off already increases the fraction of proteins for which a confident structure prediction is possible to 10–25% [53•] (Table 1). Secondary structure-based threading allows

Table 3

Some recently identified or significantly expanded protein superfamilies.

Superfamily	Enzymes with known 3D structures (PDB codes)	Enzymes with newly predicted properties	Representatives in complete genomes*	References
BRCT (conserved domain in cell cycle checkpoint proteins)	None	DNA polymerase subunit DPB11, terminal deoxynucleotidyltransferases, deoxycytidyl transferase, DNA ligases III and IV, poly(ADP-ribose) polymerase	e-1, h-1, u-1, b-1, g-1, p-1, l-1, c-2, m-0, t-0, f-0, y-9, w-8	[50]
Urease-related metal-dependent hydrolases	Urease (2kauC), phosphotriesterase (1pta), adenosine deaminase (1fkx)	AMP deaminase, adenine deaminase, cytosine deaminase, hydantoinase, dihydroorotase, allantoinase, aminoacylase, imidazolonepropionase, arylphosphatase, chlorohydrolase, formylmethanofuran dehydrogenase	e-13, h-4, u-2, b-6, g-1, p-1, l-1, c-3, m-9, t-10, f-7, y-6, w-9	[55]
Acid phosphatases	Vanadium-containing chloroperoxidase (1vnc)	Phosphatidic acid phosphatase, phosphatidylglycerol phosphatase, diacylglycerol pyrophosphate phosphatase, glucose-6-phosphatase	e-4, h-1, u-3, b-2, g-0, p-0, l-0, c-1, m-3, t-1, f-0, y-7, w-4	[56,57]
ATP-grasp (ATP-dependent C-N and C-S ligases)	Glutathione synthetase (1gsh), D-ala-D-ala ligase (2dln), biotin carboxylase (1bnc), carbamoyl phosphate synthase (1jdb), succinyl-CoA synthetase (1scu)	Phosphoribosylamine-glycine ligase, phosphoribosylglycinamide formyltransferase, phosphoribosylaminoimidazole carboxylase, tubulin-tyrosine ligase, protein S6-glutamate ligase (RimK), malate thiokinase, ATP-citrate lyase	e-10, h-5,u-4, b-12, g-2,p-2, l-1, c-7, m-9, t-8, f-11, y-10, w-15	[58]
HAD (phosphatases and other hydrolases)	L-Haloacid dehalogenase (1jud)	Phosphoserine phosphatase, phosphoglycolate phosphatase, histidinol phosphatase, glycerol-3-phosphatase, sucrose phosphate synthase, phosphomannomutase, P-type cation-transport ATPases	e-10, h-3, u-1, b-11, g-4, p-5, l-2, c-8, m-3, t-4, f-3, y-9, w-6	[52]
DHH (hydrolases)	None	Exopolyphosphatase, 5′-3′ exonuclease	e-1, h-1, u-4, b-6, g-2, p-2, l-2, c-3, m-6, t-3, f-7, y-1, w-0	[59]
Alkaline phosphatase-related metal-dependent hydrolases	Alkaline phosphatase (1alk), N-acetylgalactosamine 4-sulfatase (1fsu), cerebroside sulfatase(1auk)	Phosphopentomutase, 2,3-bisphosphoglycerate-independent phosphoglycerate mutase, streptomycin-6-phosphatase, phosphonoacetate hydrolase, phosphoglycerol transferase, nucleotide pyrophosphatase, steroid sulfatase, aryl- and hexosamine sulfatases	e-15, h-5, u-4, b-8, g-1, p-1, l-0, c-1, m-2, t-2, f-3, y-6, w-18	[40,60]

*Identified by PSI-BLAST [47••] searches of the complete genomes using the conserved motif(s) for each superfamily as a query. Organism abbreviations are as in Figure 1: e, *E. coli*; h, *H. influenzae*; u, *H. pylori*; b, *B. subtilis*; g, *M. genitalium*; p, *M. pneumoniae*; l, *B. burgdorferi*; c, *Synechocystis sp.*; m, *M. jannaschii*; t, *M. thermoautotrophicum*; f, *A. fulgidus*; y, *S. cerevisiae*; w, *C. elegans*.

another relatively small but notable increase in the predictive power [54•] (Table 1). It appears, however, that at this time, the most realistic way to further structure prediction at genome scale is to perform a complete analysis of protein superfamilies as exemplified in Table 3.

Perspective

As far as prokaryotic genomes are concerned, we have already entered the post-genomic era. While surprises certainly wait ahead, there is little doubt that the major protein families are already known or can be deciphered from the available sequences. We have recently seen major progress in methods and procedures for advanced sequence analysis, and a lot of valuable information has been extracted from the genomes. We believe, however, that a major focused effort in genome comparison is still required in order to construct a proper classification of protein families and superfamilies and systematically apply it to the goals of structural and functional prediction. Such an effort will have the potential of creating a basis for a rationally designed, decisive onslaught on structure determination and experimental identification of gene functions using computer predictions as a guide. Hopefully, this research program turns out to be both realistic and efficient.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM *et al.*: **Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd.** *Science* 1995, 269:496-512.

2. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, . Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM *et al.*: **The minimal gene complement of *Mycoplasma genitalium*.** *Science* 1995, 270:397-403.

3. Himmelreich R, Hilbert H, Plagens H, Pirkl E, Li BC, Herrmann R: **Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*.** *Nucleic Acids Res* 1996, 24:4420-4449.

4. Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG, Blake JA, FitzGerald LM, Clayton RA, Gocayne JD *et al.*: **Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*.** *Science* 1996, 273:1058-1073.

5. Kaneko T, Sato S, Kotani H, Tanaka A, Asamizu E, Nakamura Y, Miyajima N, Hirosawa M, Sugiura M, Sasamoto S *et al.*: **Sequence analysis of the genome of the unicellular *Cyanobacterium synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions.** *DNA Res* 1996, 3:109-136.

6. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M *et al.*: **Life with 6000 genes.** *Science* 1996, 274:546, 563-567.

7. •• Blattner FR, Plunkett G, 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF *et al.*: **The complete genome sequence of *Escherichia coli* K-12.** *Science* 1997, 277:1453-1474.

The completion of the genome sequence of *E. coli*, one of the classic objects of molecular biology and genetics, certainly has a symbolic significance. More importantly, the enormous amount of information available regarding *E. coli* gene functions can now be used to full potential for inferring functions of homologs in other species. However, the functions of about one half of the *E. coli* genes have not been determined experimentally, and so there is still a lot to learn about *E. coli* itself.

8. •• Tomb J, White O, Kerlavage A, Clayton R, Sutton G, Fleishmann R, Ketchum K, Klenk HP, Gill S, Dougherty BA *et al.*: **The complete genome sequence of the gastric pathogen *Helicobacter pylori*.** *Nature* 1997, 388:539-547.

The genome sequence of this bacterium is of special interest from several points of view. The genome analysis will have important practical implications as *H. pylori* is the causative agent of peptic ulcers and is believed to infect up to half of the human population. *H. pylori* thrives in a highly acidic environment (pH 2–3); deciphering the mechanisms of acid tolerance from the genome sequence is a most interesting task. Furthermore, *H. pylori* represents an early branching of the proteobacterial lineage, and the comparison of its genome with those of other Proteobacteria such as *E. coli* and *Haemophilus influenzae* will shed light on the evolution of cellular functions in bacteria and mitochondria.

9. •• Smith DR, Doucette-Stamm LA, Deloughery C, Lee H, Dubois J, Aldredge T, Bashirzadeh R, Blakely D, Cook R, Gilbert K *et al.*: **Complete genome sequence of *Methanobacterium thermoautotrophicum* deltaH: functional analysis and comparative genomics.** *J Bacteriol* 1997, 179:7135-7155.

This second genome of a methanogenic archeon to be sequenced, after *Methanococcus jannaschii*, is of major importance in corroborating trends revealed by the *M. jannaschii* genome analysis [4,34]. Like *M. jannaschii*, there is a sharp divide between the majority of the genes, which appear to have bacterial origin, and a minority (primarily encoding proteins involved in genome replication and expression) of 'eukaryotic' genes. Some other unusual aspects of the *M. jannaschii* genome, however, did not recur in *M. thermoautrophicum*. For example, unlike *M. jannaschii*, *M. thermoautrophicum* encodes a typical set of molecular chaperones such as DnaK and DnaJ and does not encode a unique ATPase family found in *M. jannaschii*.

10. •• Klenk HP, Clayton RA, Tomb JF, White O, Nelson KE, Ketchum KA, Dodson RJ, Gwinn M, Hickey EK, Peterson JD *et al.*: **The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*.** *Nature* 1997, 390:364-370.

The first sequence of a non-methanogenic archeon, and the third complete archeal genome altogether. With 2436 genes, the *A. fulgidus* genome is considerably larger than those of *M. jannaschii* and *M. thermoautrophicum*, in part due to more extensive duplication in some of the gene families. Unlike *M. jannaschii* and *M. thermoautrophicum*, *A. fulgidus* does not seem to encode any inteins. With three genome sequences available, there is for the first time an opportunity for an informative comparative analysis of archeal genomes. Definitive work in this area remains to be done, but it is already clear that the three genomes generally are highly coherent, and also that there are many mysterious conserved families, creating a challenge for further research, both theoretical and experimental.

11. •• Kunst F, Ogasawara N, Moszer I, Albertini AM, Alloni G, Azevedo V, Bertero MG, Bessieres P, Bolotin A, Borchert S *et al.*: **The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*.** *Nature* 1997, 390:249-256.

The second classic bacterial model, after *E. coli*, and also the second largest bacterial genome sequenced so far (4100 genes compared with 4288 genes in *E. coli*). With *B. subtilis* adequately representing the Gram-positive lineage (only the minimal genomes of *Mycoplasma* had been available before), we may now have a sampling of the great majority of bacterial gene families. In addition to its value for comparative analysis, *B. subtilis* is most interesting and important in its own right, given, for example the large number of genes in its genome that encode enzymes of secondary metabolite synthesis.

12. •• Fraser CM, Casjens S, Huang WM, Sutton GG, Clayton R, Lathigra R, White O, Ketchum KA, Dodson R, Hickey EK *et al.*: **Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*.** *Nature* 1997, 390:580-586.

The first genome representing yet another major division of bacteria, the spirochetes. The genome has a number of unique features, above all a linear chromosome unusual in the bacterial world, and at least 17 linear and circular plasmids that contain about 30% of the genes. Most of the plasmid-borne genes remain quite mysterious, at least after the initial genome analysis.

13. Mushegian AR, Koonin EV: **A minimal gene set for cellular life derived by comparison of complete bacterial genomes.** *Proc Natl Acad Sci USA* 1996, 93:10268-10273.

14. Koonin EV, Galperin MY: **Prokaryotic genomes: the emerging paradigm of genome-based microbiology.** *Curr Opin Genet Dev* 1997, 7:757-763.

15. Benson DA, Boguski MS, Lipman DJ, Ostell J, Ouellette BFF: **GenBank.** *Nucleic Acids Res* 1998, 26:1-7.

16. Dayhoff MO, Barker WC, Hunt LT: **Establishing homologies in protein sequences.** *Methods Enzymol* 1983, 91:524-545.

17. Barker WC, Garavelli JS, Haft DH, Hunt LT, Marzec CR, Orcutt BC, Srinivasarao GY, Yeh LSL, Ledley RS, Mewes HW *et al.*: **The PIR-**

International Protein Sequence Database. *Nucleic Acids Res* 1998, 26:27-32.

18. Bairoch A, Bucher P, Hofmann K: **The PROSITE database, its status in 1997.** *Nucleic Acids Res* 1997, 25:217-221.

19. Attwood TK, Beck ME, Flower DR, Scordis P, Selley JN: **The PRINTS protein fingerprint database in its fifth year.** *Nucleic Acids Res* 1998, 26:306-311.

20. Sonnhammer ELL, Eddy SR, Birney E, Bateman A, Durbin R: **Pfam: multiple sequence alignments and HMM-profiles of protein domains.** *Nucleic Acids Res* 1998, 26:322-325.

21. Corpet F, Gouzy J, Kahn D: **The ProDom database of protein domain families.** *Nucleic Acids Res* 1998, 26:325-328.

22. Holm L, Sander C: **Touring protein fold space with Dali/FSSP.** *Nucleic Acids Res* 1998, 26:318-321.

23. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM: **CATH – a hierarchic classification of protein domain structures.** *Structure* 1997, 5:1093-1108.

24. Hubbard TJP, Murzin AG, Brenner SE, Chothia C: **SCOP: a structural classification of proteins database.** *Nucleic Acids Res* 1997, 25:236-239.

25. Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP: a structural classification of proteins database for the investigation of sequences and structures.** *J Mol Biol* 1995, 247:536-540.

26. Fitch WM: **Distinguishing homologous from analogous proteins.** *Syst Zool* 1970, 19:99-113.

27. Fitch WM: **Uses for evolutionary trees.** *Phil Trans R Soc Lond B Biol Sci* 1995, 349:93-102.

28. Koonin EV, Tatusov RL, Rudd KE: **Sequence similarity analysis of Escherichia coli proteins: functional and evolutionary implications.** *Proc Natl Acad Sci USA* 1995, 92:11921-11925.

29. Labedan B, Riley M: **Widespread protein sequence similarities: origins of Escherichia coli genes.** *J Bacteriol* 1995, 177:1585-1588.

30. Labedan B, Riley M: **Gene products of Escherichia coli: sequence comparisons and common ancestries.** *Mol Biol Evol* 1995, 12:980-987.

31. Riley M, Labedan B: **Protein evolution viewed through Escherichia coli protein sequences: introducing the notion of a structural segment of homology, the module.** *J Mol Biol* 1997, 268:857-868.

32. Brenner SE, Hubbard T, Murzin A, Chothia C: **Gene duplications in *H. influenzae*.** *Nature* 1995, 378:140.

33. Tatusov RL, Mushegian AR, Bork P, Brown NP, Hayes WS, Borodovsky M, Rudd KE, Koonin EV: **Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli*.** *Curr Biol* 1996, 6:279-291.

34. Koonin EV, Mushegian AR, Galperin MY, Walker DR: **Comparison of**
• **archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea.** *Mol Microbiol* 1997, 25:619-637.

A detailed comparison of the first available archeal genome (*M. jannaschii*) with bacterial genomes produced a number of novel functional predictions and led to the conclusion that the majority of archeal genes most probably have a bacterial origin. Furthermore, generalizations started to emerge, including the nearly constant fraction of genes containing ancient conserved regions – about 70% in all genomes – and the same major superfamilies of paralogs.

35. Clayton RA, White O, Ketchum KA, Venter JC: **The first genome from the third domain of life.** *Nature* 1997, 387:459-462.

36. Overbeek R, Larsen N, Smith W, Maltsev N, Selkov E: **Representation of function: the next step.** *Gene* 1997, 191:GC1-GC9.

37. Tatusov RL, Koonin EV, Lipman DJ: **A genomic perspective on**
•• **protein families.** *Science* 1997, 278:631-637.

Comparative analysis of the proteins encoded in seven complete genomes from five major phylogenetic lineages and elucidation of consistent patterns of sequence similarities resulted in the delineation of 720 clusters of orthologous groups (COGs). Each COG consists of individual orthologous proteins or orthologous sets of paralogs from at least three lineages. Orthologs typically have the same function, allowing transfer of functional information from one member to an entire COG. This automatically makes possible a number of functional predictions, especially for poorly characterized genomes. The evolving system of COGs comprises a framework for functional and evolutionary genome analysis; it is accessible through the World Wide Web (http://ncbi.nlm.nih.gov/COG).

38. Himmelreich R, Plagens H, Hilbert H, Reiner B, Herrmann R: **Comparative analysis of the genomes of the bacteria *Mycoplasma pneumoniae* and *Mycoplasma genitalium*.** *Nucleic Acids Res* 1997, 25:701-712.

39. Koonin EV, Mushegian AR, Bork P: **Non-orthologous gene displacement.** *Trends Genet* 1996, 12:334-336.

40. Galperin MY, Bairoch A, Koonin EV: **A superfamily of metalloenzymes unifies phosphopentomutase and cofactor-independent phosphoglycerate mutase with alkaline phosphatases and sulfatases.** *Protein Sci* 1998, 7:in press.

41. Danson MJ: **Central metabolism of the archaea.** In *The Biochemistry of Archaea (Archaebacteria)*. Edited by Kates M, Kushner DJ, Matheson AT. Amsterdam: Elsevier; 1993:1-24.

42. Romano AH, Conway T: **Evolution of carbohydrate metabolic pathways.** *Res Microbiol* 1996, 147:448-455.

43. Bork P, Koonin EV: **Protein sequence motifs.** *Curr Opin Struct Biol* 1996, 6:366-376.

44. Bork P, Gibson TJ: **Applying motif and profile searches.** *Methods Enzymol* 1996, 266:162-184.

45. Henikoff S, Henikoff JG: **Embedding strategies for effective use of information from multiple sequence alignments.** *Protein Sci* 1997, 6:698-705.

46. Neuwald AF, Liu JS, Lipman DJ, Lawrence CE: **Extracting protein alignment models from the sequence database.** *Nucleic Acids Res* 1997, 25:1665-1677.

47. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zheng Z, Miller W,
•• Lipman DJ: **Gapped BLAST and PSI-BLAST - A new generation of protein database search programs.** *Nucleic Acids Res* 1997, 25:3389-3402.

A major revamp of BLAST, which is definitely the most popular current method for database search. The key innovations are: first, the program now makes gapped alignments, with appropriately modified statistics, which results in significant increase of sensitivity; and second, the associated program PSI (Position-Specific Iterating)-BLAST makes a position-specific weight matrix (profile) out of the first pass results and iterates searches with this profile until no new sequences with similarity scores above a defined cut-off are detected. This appears to be the most powerful existing method for detection of subtle similarities between protein sequences and delineation of protein superfamilies.

48. Mushegian AR, Bassett DE Jr, Boguski MS, Bork P, Koonin EV:
• **Positionally cloned human disease genes: patterns of evolutionary conservation and functional motifs.** *Proc Natl Acad Sci USA* 1997, 94:5831-5836.

Sequence analysis of the proteins encoded by 70 positionally cloned human disease genes showed that most of them have orthologs with the same domain architecture in the nematode, but domain rearrangements are prevalent in yeast and bacterial homologs. This is one of the first demonstrations of the utility of PSI-BLAST for the delineation of large protein superfamilies. In particular, this method was used for the identification of a conserved ATPase domain present in the repair protein MutL (one of the colon cancer gene products in humans), histidine kinases, molecular chaperones of the HSP90 family and type II DNA topoisomerases; the 3D structure for the latter was already available, defining the fold for the whole superfamily.

49. Bork P, Koonin EV: **Predicting functions from protein sequences:**
• **where are the bottlenecks?** *Nature Genet* 1998, 18:313-318.

An attempt to analyze the reasons why it is so common that functionally and phylogenetically important relationships between sequences are not detected in original analysis (particularly in the framework of genome projects) but are readily identified in subsequent, more detailed studies. It appears that the major bottlenecks include inadequate filtering for noise in sequence data (for example low-complexity sequences and very common domains) and insufficient cross-talk between different types of information.

50. Bork P, Hofmann K, Bucher P, Neuwald AF, Altschul SF, Koonin EV: **A**
•• **superfamily of conserved domains in DNA damage-responsive cell cycle checkpoint proteins.** *FASEB J* 1997, 11:68-76.

A complete description of the BRCT domain that had been originally found in BRCA1 protein and several other proteins implicated in cell cycle checkpoint. In this work, the superfamily has been extended to include a distinct version of the BRCT domain detected in bacterial DNA ligases, the large subunits of eukaryotic replication factor C, and poly(ADP-ribose) polymerases. The expansion of the BRCT domain in eukaryotes may be one of the key events in the evolution of cell-cycle control.

51. Callebaut I, Mornon JP: **From BRCA1 to RAP1: a widespread BRCT module closely associated with DNA repair.** *FEBS Lett* 1997, 400:25-30.

52. Aravind L, Galperin MY, Koonin EV: **The catalytic domain of the P-**
• **type ATPase has the haloacid dehalogenase fold.** *Trends Biochem Sci* 1998, 23:127-129.

This paper is an example of the application of sequence profile analysis to the prediction of the 3D fold and the catalytic residues in a critically important enzyme, P-ATPase, which has defied crystallization attempts and remained poorly characterized in spite of intense effort.

53. Frishman D, Mewes HW: **PEDANTic genome analysis.** *Trends Genet*
• 1997, 13:415-416.

This paper describes a very convenient Worldwide Web site compiling results of automatic analysis of all available complete genomes. The Pedant WWW site (http://pedant.mips.biochem.mpg.de/frishman/pedant.html) is arguably one of the best entry points to comparative genomics but it has to be kept in mind that it is only the first level, crude analysis that is presented here.

54. Fischer D, Eisenberg D: **Assigning folds to the proteins encoded by**
• **the genome of *Mycoplasma genitalium*.** *Proc Natl Acad Sci USA* 1997, 94:11929-11934.

One of the first systematic attempts to predict the 3D structures of proteins starting from a complete genome. The utility of sequence–structure threading is demonstrated but it also becomes clear that such methods at best result in a rather small, incremental improvement over state-of-the-art sequence comparisons. Although the fraction of the proteins with a predictable fold is only 22% of the gene products, the authors predict by extrapolation that it should be possible to assign folds to most soluble proteins within a decade.

55. Holm L, Sander C: **An evolutionary treasure: unification of a broad**
• **set of amidohydrolases related to urease.** *Proteins* 1997, 28:72-82.

A valuable example of a combination of detailed sequence analysis with structure–structure comparisons resulting in the characterization of a vast protein superfamily.

56. Stukey J, Carman GM: **Identification of a novel phosphatase sequence motif.** *Protein Sci* 1997, 6:469-472.

57. Neuwald AF: **An unexpected structural relationship between integral membrane phosphatases and soluble haloperoxidases.** *Protein Sci* 1997, 6:1764-1767.

58. Galperin MY, Koonin EV: **A diverse superfamily of enzymes with ATP-dependent carboxylate-amine/thiol ligase activity.** *Protein Sci* 1997, 6:2639-2643.

59. Aravind L, Koonin EV: **A novel family of predicted phosphoesterases includes *Drosophila* prune protein and bacterial RecJ exonuclease.** *Trends Biochem Sci* 1998, 23:17-19.

60. Bond CS, Clements PR, Ashby SJ, Collyer CA, Harrop SJ, Hopwood JJ, Guss JM: **Structure of a human lysosomal sulfatase.** *Structure* 1997, 5:277-289.

APPENDIX D

NCBI Protein

My NCBI [Sign In] [Register]

PubMed Nucleotide Protein Genome Structure PMC Taxonomy OMIM Books

Search Protein for Go Clear

Limits Preview/Index History Clipboard Details

Display GenPept Show 5 Send to

Range: from begin to end Features: ☑ CDD + Refresh

□ 1: CAF19551. Reports 3'-Phosphoadenosi...[gi:41325070]

BLink, Conserved Domains, Links

Comment Features Sequence

```
LOCUS       CAF19551                 252 aa            linear   BCT 17-APR-2005
DEFINITION  3'-Phosphoadenosine 5'-phosphosulfate (PAPS) 3'-phosphatase
            [Corynebacterium glutamicum ATCC 13032].
ACCESSION   CAF19551
VERSION     CAF19551.1  GI:41325070
DBSOURCE    embl accession BX927150.1
KEYWORDS    .
SOURCE      Corynebacterium glutamicum ATCC 13032
  ORGANISM  Corynebacterium glutamicum ATCC 13032
            Bacteria; Actinobacteria; Actinobacteridae; Actinomycetales;
            Corynebacterineae; Corynebacteriaceae; Corynebacterium.
REFERENCE   1  (residues 1 to 252)
  AUTHORS   Kalinowski,J., Bathe,B., Bartels,D., Bischoff,N., Bott,M.,
            Burkovski,A., Dusch,N., Eggeling,L., Eikmanns,B.J., Gaigalat,L.,
            Goesmann,A., Hartmann,M., Huthmacher,K., Kramer,R., Linke,B.,
            McHardy,A.C., Meyer,F., Mockel,B., Pfefferle,W., Puhler,A.,
            Rey,D.A., Ruckert,C., Rupp,O., Sahm,H., Wendisch,V.F., Wiegrabe,I.
            and Tauch,A.
  TITLE     The complete Corynebacterium glutamicum ATCC 13032 genome sequence
            and its impact on the production of L-aspartate-derived amino acids
            and vitamins
  JOURNAL   J. Biotechnol. 104 (1-3), 5-25 (2003)
   PUBMED   12948626
REFERENCE   2  (residues 1 to 252)
  AUTHORS   Kalinowski,J.
  TITLE     Direct Submission
  JOURNAL   Submitted (21-JAN-2004) Joern Kalinowski, Institut fuer
            Genomforschung, Universitaet Bielefeld; Universitaetsstrasse 25,
            33615 Bielefeld, Germany
            E-mail:Joern.Kalinowski@Cebitec.Uni-Bielefeld.DE
COMMENT     This sequence was accomplished by collaboration between Degussa AG
            and Bielefeld University.
            join(BX927148.1:1..348071,BX927149.1:51..349887,
            BX927150.1:51..348475,
            BX927151.1:51..349459,BX927152.1:51..349799,BX927153.1:51..349584,
            BX927154.1:51..349575,BX927155.1:51..349136,BX927156.1:51..349115,
            BX927157.1:51..140057).
FEATURES             Location/Qualifiers
     source          1..252
                     /organism="Corynebacterium glutamicum ATCC 13032"
                     /strain="DSM 20300 = ATCC 13032"
                     /db_xref="taxon:196627"
                     /note="IS fingerprint type: 4-5"
     Protein         1..252
                     /product="3'-Phosphoadenosine 5'-phosphosulfate (PAPS)
                     3'-phosphatase"
     Region          10..>229
                     /region_name="CysQ, a
                     3'-Phosphoadenosine-5'-phosphosulfate (PAPS)
                     3'-phosphatase, is a bacterial member of the inositol
                     monophosphatase family"
                     /note="CysQ"
                     /db_xref="CDD:30136"
     CDS             1..252
                     /gene="cysQ"
                     /locus_tag="cg0967"
                     /coded_by="complement(BX927150.1:202863..203621)"
                     /transl_table=11
                     /db_xref="GOA:Q8NS37"
                     /db_xref="InterPro:IPR000760"
                     /db_xref="UniProtKB/TrEMBL:Q8NS37"
ORIGIN
        1 mtaqiddsil thrlaqgtge ilkgvrnvgv lrgrnlgdag delaqswiar vleqhrpndg
       61 flseeaadnp drlskdrvwi idpldgtkef atgrqdwavh ialvengvpt haavglpdlg
      121 vvfhsadara vtgpyskvia ishnrppkva lscaeqlgfe tkalgsagak amhvllgdyd
      181 ayihaggqye wdsaapvgvc kaaglhcsrl dgseltynnk dtympdilic rpeladelle
      241 mcakfyeeng ty
//
```

Disclaimer | Write to the Help Desk
NCBI | NLM | NIH

Apr 11 2006 19:57:30

APPENDIX E

NCBI Entrez Protein

My NCBI [Sign In] [Register]

PubMed Nucleotide Protein Genome Structure PMC Taxonomy OMIM Books

Search Protein for Go Clear

Limits Preview/Index History Clipboard Details

Display GenPept Show 5 Send to

Range: from begin to end Features: ☑ CDD + Refresh

□ 1: BAB98238. Reports 3'-Phosphoadenosi...[gi:21323611]

BLink, Conserved Domains, Links

Comment Features Sequence

```
LOCUS       BAB98238                 252 aa            linear   BCT 03-FEB-2005
DEFINITION  3'-Phosphoadenosine 5'-phosphosulfate (PAPS) 3'-phosphatase
            [Corynebacterium glutamicum ATCC 13032].
ACCESSION   BAB98238
VERSION     BAB98238.1  GI:21323611
DBSOURCE    accession BA000036.3
KEYWORDS    .
SOURCE      Corynebacterium glutamicum ATCC 13032
  ORGANISM  Corynebacterium glutamicum ATCC 13032
            Bacteria; Actinobacteria; Actinobacteridae; Actinomycetales;
            Corynebacterineae; Corynebacteriaceae; Corynebacterium.
REFERENCE   1
  AUTHORS   Nakagawa,S.
  TITLE     Complete genomic sequence of Corynebacterium glutamicum ATCC 13032
  JOURNAL   Unpublished
REFERENCE   2  (residues 1 to 252)
  AUTHORS   Nakagawa,S.
  TITLE     Direct Submission
  JOURNAL   Submitted (24-MAY-2002) Satoshi Nakagawa, Kyowa Hakko Kogyo Co.
            Ltd., Tokyo Research Laboratories; 3-6-6, Asahi-machi, Machida,
            Tokyo, 194-8533, Japan (E-mail:snakagawa@xanagen.com,
            Tel:81-44-829-3031, Fax:81-44-813-1651)
COMMENT     This sequence is conducted by collaboration of Kyowa Hakko Kogyo
            Co. Ltd. And Kitasato University.
FEATURES             Location/Qualifiers
     source          1..252
                     /organism="Corynebacterium glutamicum ATCC 13032"
                     /strain="ATCC 13032"
                     /db_xref="taxon:196627"
     Protein         1..252
                     /product="3'-Phosphoadenosine 5'-phosphosulfate (PAPS)
                     3'-phosphatase"
     Region          10..>229
                     /region_name="CysQ, a
                     3'-Phosphoadenosine-5'-phosphosulfate (PAPS)
                     3'-phosphatase, is a bacterial member of the inositol
                     monophosphatase family"
                     /note="CysQ"
                     /db_xref="CDD:30136"
     CDS             1..252
                     /gene="Cgl0845"
                     /coded_by="complement(BA000036.2:899250..900008)"
                     /note="PF00459:Inositol monophosphatase family"
                     /transl_table=11
ORIGIN
        1 mtaqiddsil thrlaqgtge ilkgvrnvgv lrgrnlgdag delaqswiar vleqhrpndg
       61 flseeaadnp drlskdrvwi idpldgtkef atgrqdwavh ialvengvpt haavglpdlg
      121 vvfhsadara vtgpyskvia ishnrppkva lscaeqlgfe tkalgsagak amhvllgdyd
      181 ayihaggqye wdsaapvgvc kaaglhcsrl dgseltynnk dtympdilic rpeladelle
      241 mcakfyeeng ty
//
```

Disclaimer | Write to the Help Desk
NCBI | NLM | NIH

Apr 11 2006 19:57:30

APPENDIX F

My NCBI
[Sign In]
[Register]

PubMed Nucleotide Protein Genome Structure PMC Taxonomy OMIM Books

Search Protein for Go Clear

Limits Preview/Index History Clipboard Details

Display GenPept Show 5 Send to

Range: from begin to end Features: ☑ CDD + Refresh

□ 1: YP_225137. Reports 3'-Phosphoadenosi...[gi:62389735]

BLink, Conserved Domains, Links

Comment Features Sequence

```
LOCUS       YP_225137                252 aa            linear   BCT 17-JAN-2006
DEFINITION  3'-Phosphoadenosine 5'-phosphosulfate (PAPS) 3'-phosphatase
            [Corynebacterium glutamicum ATCC 13032].
ACCESSION   YP_225137
VERSION     YP_225137.1  GI:62389735
DBSOURCE    REFSEQ: accession NC_006958.1
KEYWORDS    complete genome.
SOURCE      Corynebacterium glutamicum ATCC 13032
  ORGANISM  Corynebacterium glutamicum ATCC 13032
            Bacteria; Actinobacteria; Actinobacteridae; Actinomycetales;
            Corynebacterineae; Corynebacteriaceae; Corynebacterium.
REFERENCE   1  (residues 1 to 252)
  AUTHORS   Kalinowski,J., Bathe,B., Bartels,D., Bischoff,N., Bott,M.,
            Burkovski,A., Dusch,N., Eggeling,L., Eikmanns,B.J., Gaigalat,L.,
            Goesmann,A., Hartmann,M., Huthmacher,K., Kramer,R., Linke,B.,
            McHardy,A.C., Meyer,F., Mockel,B., Pfefferle,W., Puhler,A.,
            Rey,D.A., Ruckert,C., Rupp,O., Sahm,H., Wendisch,V.F., Wiegrabe,I.
            and Tauch,A.
  TITLE     The complete Corynebacterium glutamicum ATCC 13032 genome sequence
            and its impact on the production of L-aspartate-derived amino acids
            and vitamins
  JOURNAL   J. Biotechnol. 104 (1-3), 5-25 (2003)
   PUBMED   12948626
REFERENCE   2  (residues 1 to 252)
  CONSRTM   NCBI Genome Project
  TITLE     Direct Submission
  JOURNAL   Submitted (07-APR-2005) National Center for Biotechnology
            Information, NIH, Bethesda, MD 20894, USA
REFERENCE   3  (residues 1 to 252)
  AUTHORS   Kalinowski,J.
  TITLE     Direct Submission
  JOURNAL   Submitted (21-JAN-2004) Institut fuer Genomforschung, Universitaet
            Bielefeld, Universitaetsstrasse 25, Bielefeld 33615, Germany
COMMENT     PROVISIONAL REFSEQ: This record has not yet been subject to final
            NCBI review. The reference sequence was derived from CAF19551.
            Method: conceptual translation.
FEATURES             Location/Qualifiers
     source          1..252
                     /organism="Corynebacterium glutamicum ATCC 13032"
                     /strain="DSM 20300; ATCC 13032"
                     /db_xref="ATCC:13032"
                     /db_xref="taxon:196627"
                     /note="IS fingerprint type 4-5"
     Protein         1..252
                     /product="3'-Phosphoadenosine 5'-phosphosulfate (PAPS)
                     3'-phosphatase"
                     /calculated_mol_wt=27151
     Region          10..>229
                     /region_name="CysQ, a
                     3'-Phosphoadenosine-5'-phosphosulfate (PAPS)
                     3'-phosphatase, is a bacterial member of the inositol
                     monophosphatase family"
                     /note="CysQ"
                     /db_xref="CDD:30136"
     CDS             1..252
                     /gene="cysQ"
                     /locus_tag="cg0967"
                     /coded_by="complement(NC_006958.1:900721..901479)"
                     /transl_table=11
                     /db_xref="GeneID:3345270"
ORIGIN
        1 mtaqiddsil thrlaqgtge ilkgvrnvgv lrgrnlgdag delaqswiar vleqhrpndg
       61 flseeaadnp drlskdrvwi idpldgtkef atgrqdwavh ialvengvpt haavglpdlg
      121 vvfhsadara vtgpyskvia ishnrppkva lscaeqlgfe tkalgsagak amhvllgdyd
      181 ayihaggqye wdsaapvgvc kaaglhcsrl dgseltynnk dtympdilic rpeladelle
      241 mcakfyeeng ty
//
```

Disclaimer | Write to the Help Desk
NCBI | NLM | NIH

Apr 11 2006 19:57:30

APPENDIX G

lalign output for SEQ ID NO:6 vs. CAF19551

[ISREC-Server] Date: Mon Jun 26 19:28:01 Europe/Zurich 2006

./wwwtmp/lalign/.19436.1.seq : 252 aa

```
ALIGN calculates a global alignment of two sequences
 version 2.0uPlease cite: Myers and Miller, CABIOS (1989) 4:11-17
SEQ ID NO:6                                         252 aa vs.
CAF19551                                            252 aa
scoring matrix: BLOSUM50, gap penalties: -14/-4
100.0% identity;                Global alignment score: 1703

               10        20        30        40        50        60
./wwwt MTAQIDDSILTHRLAQGTGEILKGVRNVGVLRGRNLGDAGDELAQSWIARVLEQHRPNDG
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
CAF195 MTAQIDDSILTHRLAQGTGEILKGVRNVGVLRGRNLGDAGDELAQSWIARVLEQHRPNDG
               10        20        30        40        50        60

               70        80        90       100       110       120
./wwwt FLSEEAADNPDRLSKDRVWIIDPLDGTKEFATGRQDWAVHIALVENGVPTHAAVGLPDLG
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
CAF195 FLSEEAADNPDRLSKDRVWIIDPLDGTKEFATGRQDWAVHIALVENGVPTHAAVGLPDLG
               70        80        90       100       110       120

              130       140       150       160       170       180
./wwwt VVFHSADARAVTGPYSKVIAISHNRPPKVALSCAEQLGFETKALGSAGAKAMHVLLGDYD
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
CAF195 VVFHSADARAVTGPYSKVIAISHNRPPKVALSCAEQLGFETKALGSAGAKAMHVLLGDYD
              130       140       150       160       170       180

              190       200       210       220       230       240
./wwwt AYIHAGGQYEWDSAAPVGVCKAAGLHCSRLDGSELTYNNKDTYMPDILICRPELADELLE
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
CAF195 AYIHAGGQYEWDSAAPVGVCKAAGLHCSRLDGSELTYNNKDTYMPDILICRPELADELLE
              190       200       210       220       230       240

              250
./wwwt MCAKFYEENGTY
       ::::::::::::
CAF195 MCAKFYEENGTY
              250
```

Back to ISREC bioinformatics group home page

APPENDIX H

lalign output for SEQ ID NO:6 vs. BAB98238

[ISREC-Server] Date: Mon Jun 26 19:35:51 Europe/Zurich 2006

./wwwtmp/lalign/.12164.1.seq : 252 aa

```
ALIGN calculates a global alignment of two sequences
 version 2.0uPlease cite: Myers and Miller, CABIOS (1989) 4:11-17
SEQ ID NO:6                                        252 aa vs.
BAB98238                                           252 aa
scoring matrix: BLOSUM50, gap penalties: -14/-4
100.0% identity;              Global alignment score: 1703

              10        20        30        40        50        60
./wwwt MTAQIDDSILTHRLAQGTGEILKGVRNVGVLRGRNLGDAGDELAQSWIARVLEQHRPNDG
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
BAB982 MTAQIDDSILTHRLAQGTGEILKGVRNVGVLRGRNLGDAGDELAQSWIARVLEQHRPNDG
              10        20        30        40        50        60

              70        80        90       100       110       120
./wwwt FLSEEAADNPDRLSKDRVWIIDPLDGTKEFATGRQDWAVHIALVENGVPTHAAVGLPDLG
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
BAB982 FLSEEAADNPDRLSKDRVWIIDPLDGTKEFATGRQDWAVHIALVENGVPTHAAVGLPDLG
              70        80        90       100       110       120

             130       140       150       160       170       180
./wwwt VVFHSADARAVTGPYSKVIAISHNRPPKVALSCAEQLGFETKALGSAGAKAMHVLLGDYD
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
BAB982 VVFHSADARAVTGPYSKVIAISHNRPPKVALSCAEQLGFETKALGSAGAKAMHVLLGDYD
             130       140       150       160       170       180

             190       200       210       220       230       240
./wwwt AYIHAGGQYEWDSAAPVGVCKAAGLHCSRLDGSELTYNNKDTYMPDILICRPELADELLE
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
BAB982 AYIHAGGQYEWDSAAPVGVCKAAGLHCSRLDGSELTYNNKDTYMPDILICRPELADELLE
             190       200       210       220       230       240

             250
./wwwt MCAKFYEENGTY
       ::::::::::::
BAB982 MCAKFYEENGTY
             250
```

Back to ISREC bioinformatics group home page

APPENDIX I

lalign output for SEQ ID NO:6 vs. YP_225137

[ISREC-Server] Date: Mon Jun 26 19:39:50 Europe/Zurich 2006

./wwwtmp/lalign/.4363.1.seq : 252 aa

```
ALIGN calculates a global alignment of two sequences
 version 2.0uPlease cite: Myers and Miller, CABIOS (1989) 4:11-17
SEQ ID NO:6                                            252 aa vs.
YP_225137                                              252 aa
scoring matrix: BLOSUM50, gap penalties: -14/-4
100.0% identity;                Global alignment score: 1703

               10        20        30        40        50        60
./wwwt MTAQIDDSILTHRLAQGTGEILKGVRNVGVLRGRNLGDAGDELAQSWIARVLEQHRPNDG
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
YP_225 MTAQIDDSILTHRLAQGTGEILKGVRNVGVLRGRNLGDAGDELAQSWIARVLEQHRPNDG
               10        20        30        40        50        60

               70        80        90       100       110       120
./wwwt FLSEEAADNPDRLSKDRVWIIDPLDGTKEFATGRQDWAVHIALVENGVPTHAAVGLPDLG
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
YP_225 FLSEEAADNPDRLSKDRVWIIDPLDGTKEFATGRQDWAVHIALVENGVPTHAAVGLPDLG
               70        80        90       100       110       120

              130       140       150       160       170       180
./wwwt VVFHSADARAVTGPYSKVIAISHNRPPKVALSCAEQLGFETKALGSAGAKAMHVLLGDYD
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
YP_225 VVFHSADARAVTGPYSKVIAISHNRPPKVALSCAEQLGFETKALGSAGAKAMHVLLGDYD
              130       140       150       160       170       180

              190       200       210       220       230       240
./wwwt AYIHAGGQYEWDSAAPVGVCKAAGLHCSRLDGSELTYNNKDTYMPDILICRPELADELLE
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
YP_225 AYIHAGGQYEWDSAAPVGVCKAAGLHCSRLDGSELTYNNKDTYMPDILICRPELADELLE
              190       200       210       220       230       240

              250
./wwwt MCAKFYEENGTY
       ::::::::::::
YP_225 MCAKFYEENGTY
              250
```

Back to ISREC bioinformatics group home page

APPENDIX J

JOURNAL OF BACTERIOLOGY, Jan. 1992, p. 415–425
0021-9193/92/020415-11$02.00/0
Copyright © 1992, American Society for Microbiology

Vol. 174, No. 2

cysQ, a Gene Needed for Cysteine Synthesis in *Escherichia coli* K-12 Only during Aerobic Growth

ANDREW F. NEUWALD,[1,2] B. RAJENDRA KRISHNAN,[1] IGOR BRIKUN,[1] SAULIUS KULAKAUSKAS,[1†] KĘSTUTIS SUŽIEDĖLIS,[1‡] TIHAMER TOMCSANYI,[1§] THOMAS S. LEYH,[3] AND DOUGLAS E. BERG[1,4*]

Department of Molecular Microbiology, Box 8230,[1] Institute for Biomedical Computing,[2] and Department of Genetics,[4] Washington University Medical School, St. Louis, Missouri 63110-1093, and Department of Biochemistry, Albert Einstein College of Medicine, Bronx, New York 10461[3]

Received 31 July 1991/Accepted 6 November 1991

The initial steps in assimilation of sulfate during cysteine biosynthesis entail sulfate uptake and sulfate activation by formation of adenosine 5′-phosphosulfate, conversion to 3′-phosphoadenosine 5′-phosphosulfate, and reduction to sulfite. Mutations in a previously uncharacterized *Escherichia coli* gene, *cysQ*, which resulted in a requirement for sulfite or cysteine, were obtained by in vivo insertion of transposons Tn*5tacl* and Tn*5supF* and by in vitro insertion of resistance gene cassettes. *cysQ* is at chromosomal position 95.7 min (kb 4517 to 4518) and is transcribed divergently from the adjacent *cpdB* gene. A Tn*5tacl* insertion just inside the 3′ end of *cysQ*, with its isopropyl-β-D-thiogalactopyranoside-inducible *tac* promoter pointed toward the *cysQ* promoter, resulted in auxotrophy only when isopropyl-β-D-thiogalactopyranoside was present; this conditional phenotype was ascribed to collision between converging RNA polymerases or interaction between complementary antisense and *cysQ* mRNAs. The auxotrophy caused by *cysQ* null mutations was leaky in some but not all *E. coli* strains and could be compensated by mutations in unlinked genes. *cysQ* mutants were prototrophic during anaerobic growth. Mutations in *cysQ* did not affect the rate of sulfate uptake or the activities of ATP sulfurylase and its protein activator, which together catalyze adenosine 5′-phosphosulfate synthesis. Some mutations that compensated for *cysQ* null alleles resulted in sulfate transport defects. *cysQ* is identical to a gene called *amtA*, which had been thought to be needed for ammonium transport. Computer analyses, detailed elsewhere, revealed significant amino acid sequence homology between *cysQ* and *suhB* of *E. coli* and the gene for mammalian inositol monophosphatase. Previous work had suggested that 3′-phosphoadenoside 5′-phosphosulfate is toxic if allowed to accumulate, and we propose that CysQ helps control the pool of 3′-phosphoadenoside 5′-phosphosulfate, or its use in sulfite synthesis.

The cysteine biosynthetic pathway (Fig. 1), a principal route of sulfur assimilation, involves more than 15 genes in at least five chromosomal regions in *Escherichia coli* and *Salmonella typhimurium*. It has been studied since the early days of physiological genetics in order to elucidate the roles of the individual genes, the control of their expression, and how the flow of metabolic intermediates is regulated (for a review, see reference 25). The transcription of most *cys* genes is positively controlled by the protein product of *cysB* and its coinducer, *O*-acetyl serine (also a cysteine precursor), during aerobic growth; transcription is repressed by sulfide, which is generated by reversal of the final biosynthetic step (Fig. 1). CysB seems not to be needed during anaerobic growth (3). The *cysQ* gene described here is also needed only during aerobic growth. It is inferred to act before sulfite formation, and hence this early part of the cysteine pathway is reviewed briefly below.

The initial step, sulfate uptake, is mediated by a permease encoded by the *cysT*, *cysW*, and *cysA* genes, which, along with *cysP*, constitute one operon (49). CysP protein is needed for maximal thiosulfate and sulfate binding, but it is probably not part of the permease, and its role in cysteine biosynthesis is unclear (19). A *cysZ* gene, about 10 kb from the *cysPTWA* operon, may also be needed for sulfate uptake (41). Intracellular sulfate is activated via synthesis of adenosine 5′-phosphosulfate (APS) by ATP sulfurylase, which is encoded by *cysD* and *cysN* (34). This activation step is complex, in that the rate of APS formation is greatly enhanced both by a protein activator (31, 32) and by GTP hydrolysis (33). APS is converted to 3′-phosphoadenosine 5′-phosphosulfate (PAPS) by APS kinase, encoded by *cysC*. This step is thought to not require cofactors, because APS kinase activity does not change during enzyme purification (46). Sulfite is generated from PAPS in a complex reaction involving transfer and reduction of its sulfuryl moiety. This reaction is catalyzed by the *cysH* gene product, PAPS sulfotransferase, and involves a thioredoxin- or glutaredoxin-bound intermediate (51, 52).

Strains with mutations in *cysH* or in both *trxA* and *grx* (encoding thioredoxin and glutaredoxin, respectively) grow poorly. The poor growth can be corrected by additional mutations in *cysC* (APS kinase) or genes for earlier steps in the pathway (16, 45a), a result indicating that PAPS or one of its derivatives is toxic if allowed to accumulate. We find this result interesting in the context of understanding mechanisms by which organisms cope with the many metabolic intermediates that are both essential for healthy growth and potentially deleterious. Other studies have shown that the activities of ATP sulfurylase and APS kinase decrease rapidly when growth is slowed (25, 26). Such instability could help modulate metabolite flow through this pathway and would be more sensitive to decreased need for PAPS

[*] Corresponding author.

[†] Permanent address: Institute of Applied Enzymology, FERMENTAS, Vilnius 232028, Lithuanian Republic.

[‡] Permanent address: Department of Biochemistry and Biophysics, Vilnius University, Vilnius 232009, Lithuanian Republic.

[§] Present address: Department of Zoology, Janus Pannonius University, Ifjusag uta 6, 7601 Pecs, Hungary.

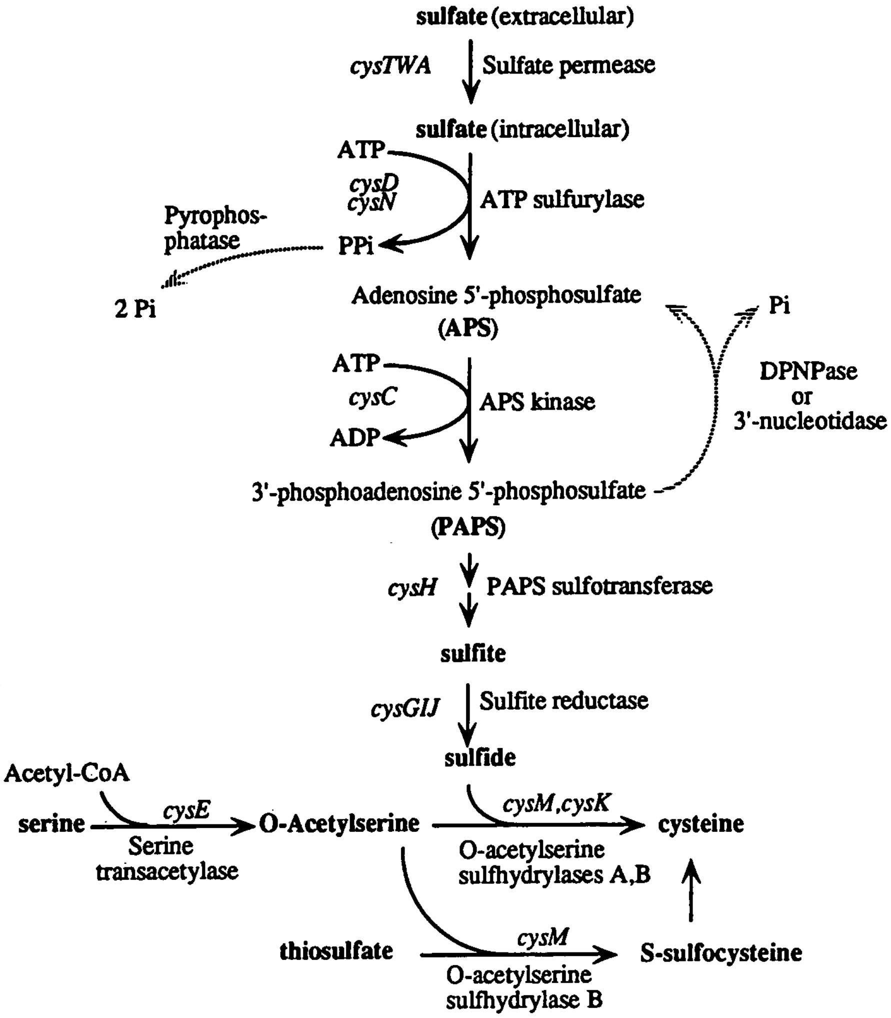


FIG. 1. Pathway of cysteine biosynthesis in *E. coli* (modified from data in references 25, 34, and 49).

than any transcriptional regulation. The dependence of APS synthesis (and thereby PAPS synthesis) on the ATP sulfurylase activator and on the local concentration of GTP (32, 33) might also help regulate PAPS levels.

Mutations in the *cysQ* gene described here result in a requirement for cysteine or sulfite that is expressed only during aerobic growth and that is leaky in many but not all laboratory strains of *E. coli*. Our studies suggest that CysQ may help control the levels of PAPS, its localization, or its use in sulfite synthesis.

MATERIALS AND METHODS

Strains, media, and general methods. The bacterial strains and plasmids used in this study are listed in Table 1. Bacteria were grown in LN broth (5) or Vogel-Bonner glucose-minimal salts medium (54). An M9-based minimal salts medium, with ammonium acetate in place of ammonium chloride (21), was used where indicated. Solid media contained 1.5% Difco Bacto-Agar. Antibiotics were used at the following concentrations: ampicillin, 250 μg/ml; kanamycin, 60 μg/ml; tetracycline, 12 μg/ml; streptomycin, 100 μg/ml; and chloramphenicol, 20 μg/ml. Isopropyl-β-D-thiogalactopyranoside (IPTG) was used at 0.5 mM. Amino acids were added at 50 μg/ml except for glycine, which was added at 200 μg/ml. Standard procedures were used for bacterial growth, characterization of auxotrophy and conjugation, DNA preparation, restriction endonuclease digestion, DNA electrophoresis, recombinant DNA cloning, and transformation (12, 47). All enzymes were obtained from commercial sources (Life Technologies, Inc., Stratagene, New England BioLabs, or Boehringer Mannheim) and used as directed. Anaërobic (H_2-CO_2 atmosphere) conditions were obtained by using BBL GasPak Anaerobic jars and the BBL GasPak plus system (Becton Dickinson and Co.).

Assays. The activity of 2',3'-cyclic phosphodiesterase (encoded by the *cpdB* gene) was measured as the release of inorganic phosphate from cyclic UMP (4). Sulfate uptake was measured as depletion of $^{35}SO_4$ added to the medium with cells grown with djenkolic acid as the sulfur source and concentrated from the exponential phase (19). ATP sufurylase was measured as incorporation of $^{35}SO_4$ into PAPS, detected by thin-layer chromatography with dialyzed extracts of cells grown with sulfite as the sulfur source and induced with *O*-acetyl-1-serine (34). The level of the activator of ATP sulfurylase was also measured as PAPS synthesis

TABLE 1. Bacterial strains, phage, and plasmids

Strain, phage, or plasmid	Description or genotype	Source or reference
E. coli		
AJ2653^a	ET8000 *amtA* (*cysQ*)::Tn*10*	21
BW6458	*proC*::Tn*5 zje*::Tn*10*-BJW43 *metB1 relA1*	B. Wanner
CAG5052	Hfr *btuB*3191::Tn*10*, transfer counterclockwise from 7 min	48
DB747	W3350 *gal rpsL sup*^{0} (strain 594 of reference 7)	Laboratory collection
DB1434	DB747 (λ*plac5 c*I857 *S*am7)	28
DB4496	MC1061 *dam*::Tn*9* (p3)(pBRG1310)	43
DB5463	HfrH *lacZ*(Am) *trp*(Am) *sup*^{0}	D. Botstein (DB6128)
DB5508	*recD* (p3)	44
DB5659	DB747 *cysQ*::*kan*	λ*cysQ*::*kan* transduction
DB6302^a	MG1655 *amtA* (*cysQ*)::Tn*10*	P1 transduction from AJ2653
DB6316	MG1655 *cysQ*::*kan*	λ*cysQ*::*kan* transduction
DB6908	ET8000 *cysQ*::*kan*	λ*cysQ*::*kan* transduction
DB6913	TG1 *cysQ*::*kan*	λ*cysQ*::*kan* transduction
DB6935	DB5508 *cysQ*::Tn*5supF*	This study
DB7101	TG1 Δ(*cysD-N*)::*kan*	λΔ(*cysD-N*)::*kan* transduction
DBan41	594 with *cysQ*::Tn*5tacI*	Tn*5tacI* transposition
DBan41TR	DBan41 with Δ (*srl-recA*)*306*::Tn*10*	P1 transduction from JC1289
DK21	*sup*^{0} *dnaB*(Am) *266* (λ*imm*^{21}-*ban*$_{P1}$)	29
ET8000	*rbs lacZ*::IS*1 gyrA hutC*$_k$	37
JC1289	Δ(*srl-recA*)*306* linked to Tn*10*	11
MC1061	F$^-$ *araD139*Δ(*ara-leu*)*7697* Δ*lacX74 galU galK hsdR hsdM rpsL*	8
MG1655	F$^-$ prototroph	17
TG1	F′ *proAB*$^+$ *traD*36 *lacI*^{q} *lacZ*ΔM15 *supE hsd*Δ5 *thi* Δ(*lac-proAB*)	47
BW6164	HfrRA2 *thr*::Tn*10*, clockwise transfer from 88 min	55
Phages		
M13mp18	Cloning vector	38
λ$^+$	λ wild type	Laboratory collection
λ::Tn*5tacI*	λTn*5tacI b*221 *c*I857 *O*am29 *P*am80	9
λ419	*cysA*$^+$ (5F7 of reference 24)	23
λ656	*cysQ*$^+$ (5B5 of reference 24)	23
λΔ(*cysD-N*)::*kan*	Derivative of λ652 (6C8 of reference 24)	28
λ*cysQ*::*kan*	Derivative of λ656	28
λ656*cysQ*::Tn*5supF*		This study
P1clr		Laboratory collection
Plasmids		
pBRGan101	Amp^r, *cysQ*::Tn*5tacI* (Kan^r) *cysQ*$^+$	pBR322 (*Sal*I fragment from DBan41)
pBRGan102	Amp^r, *cysQ*::Tn*5tacI* (Kan^r) *cysQ*$^+$ *cpdB*$^+$	pBR322 (*Cla*I fragment from DBan41)
pBRGan103	Amp^r, *cysQ*$^+$ *cpdB*$^+$ *Cla*I::Tn*5tacI* Δ(0–52)	pBRGan102, small *Cla*I deletion (Kan^r)
pBRGan104	Amp^r, *cpdB*$^+$	pBRGan102, *Eco*RI deletion
pBRGan110	Amp^r, *cysQ*$^+$	pBR322 (*Sal*I fragment of DB747)
pBRGan111	Amp^r, *cysQ*$^+$	pAN110, partial *Msp*I digestion
pBRGan111-1 to pBRGan111-5	Amp^r, *cysQ*::*cat*	*cat* of pCM4 ligate into partial *Sau*3A of pBRGan111 (Fig. 2 and 4)
p*cysQ*::*kan*	Amp^r, *cysQ*::*kan*	*kan* of pUC4K into *Eco*RI site of pBRGan111
pCM4		10
pBR322	Amp^r Tet^r	6
p3	Kan^r *amp*(Am) *tet*(Am)	29
pBRG1310	Tn*5supF* donor	43

[a] The *amtA* gene is identical to *cysQ*, as detailed in the text.

in reactions containing purified ATP sulfurylase and dialyzed cell extracts as the source of ATP sulfurylase activator (31, 32).

Genetic manipulation and analysis. Standard methods were used for (i) mutagenesis of *E. coli* with Tn*5tacI* with phage λ::Tn*5tacI* b221 *c*I857 *O*am29 *P*am80 as a transposon donor (9) and (ii) Hfr conjugation and P1 generalized transduction (48). Cysteine-requiring bacteria were tested for sensitivity to azaserine and to chromate by spotting dilutions of these agents on lawns of 10^7 bacteria spread on minimal glucose agar supplemented with cysteine, djenkolic acid, or glutathione and IPTG, as appropriate, and by growth in liquid cultures with progressive twofold differences in the concentrations of these agents.

To insert a transcription reporter into the *cysQ* gene, the DNA of *cysQ* plasmid pBRGan111 was partially digested with *Sau*3A, and full-length linear DNA was isolated after electrophoresis in low-melting-point agarose and ligated to the *Bam*HI *cat* fragment from plasmid pCM-4 (10). The religated DNA was used to transform the *cysQ*::Tn*5tacl* strain DBan41, and plasmids that did not complement its cysteine auxotrophy were identified and characterized.

To generate a *cysQ*::Tn*5supF* insertion mutant, Tn*5supF* was transposed from the donor plasmid in strain DB4496 (43) to *cysQ*$^+$ phage λ656 (23, 24), and insertion-containing phage were selected by plaque formation on the *dnaB* amber strain DK21 (29, 43). Haploid Tn*5supF*-containing bacterial recombinants were obtained by infecting strain DB5508 (which contains amber mutant alleles of *amp* and *tet* genes) and selecting Sup$^+$ transductants by their resistance to ampicillin or tetracycline (44). Sup$^+$ transductants were screened for auxotrophy. To generate a *cysQ*::*kan* mutant, an *Eco*RI *kan* cassette from plasmid pUC4K was ligated into the *Eco*RI site in *cysQ* of pBRGan111. This allele was recombined into λ656 by infecting cells carrying the pBRGan111-*cysQ*::*kan* plasmid and selecting phage carrying the *cysQ*::*kan* allele by transduction of DB1434. λ*cysQ*::*kan* phage recovered from the lysogen were used to transduce nonlysogens and thereby obtain haploid *cysQ*::*kan* bacteria (28).

Cys$^+$ revertants of *cysQ* mutant strains were obtained by growing young single-colony isolates in 2 ml of LN broth to stationary phase, washing the cells twice with 10 mM $MgSO_4$, plating aliquots on minimal (cysteine-free) medium, and incubating for 2 days at 37°C. Reversion frequencies were measured by using several cultures from different single colonies to avoid jackpots.

DNA sequence analysis. A 1-kb segment containing the *cysQ* gene was sequenced by the Sanger dideoxynucleotide-chain termination method with Sequenase (U.S. Biochemical, Cleveland, Ohio) and single- and double-stranded DNA templates (27). Primer binding sites were provided by insertions of transposons Tn*5tacl* and Tn*5supF* in phage λ656, by the promoterless *cat* gene in plasmid pBRGan111 DNAs, and by a universal primer binding site in M13mp18 (38) (for sequencing an *Eco*RI-*Pst*I fragment containing the 3′ end of *cysQ*).

The oligonucleotides used as sequencing primers are as follows: (i) 5′ CTCCATTTTAGCTTCCTTAGCTCC, positions 40 through 17 at the 5′ end of the *cat* gene cassette; (ii) 5′ TGTCAAAACATGAGAATTCCTCCCG, positions 43 through 20 near the I end of Tn*5tacl*; (iii) 5′ GGAAACAGA ATTCCCGGGGATCCCC, positions 4549 through 4573 near the O end of Tn*5tacl*; (iv) 5′ TAGGATCCCCTACTTGT GTA, positions 30 through 11 near the O end of Tn*5supF*; (v) 5′ TAGGATCCCGAGATCTGATC, positions 236 through 255 near the I end of Tn*5supF*; (vi) 5′ GAGCGGCC AAAGGGAGCAGAC, positions 139 through 159 (middle primer) within Tn*5supF* with its 3′ end toward the I end; (vii) 5′ GTAAAACGACGGCCAGT, the universal M13 sequencing primer.

Nucleotide sequence accession number. The nucleotide sequence of the *cysQ* gene shown in Fig. 4 has been deposited with GenBank under accession number M80795.

RESULTS

Initial detection and characterization of *cysQ*. The prototrophic strain *E. coli* DB747 was mutagenized with Tn*5tacl*, a transposon with an outward-facing *tac* promoter that is regulated by the *lac* repressor and IPTG (9). A conditional mutant that required cysteine for growth on minimal medium containing IPTG, but not on medium lacking IPTG, was isolated and named DBan41. Early characterizations of this strain revealed two other novel features. It did not require cysteine for normal growth in an anaerobic atmosphere. In addition, it formed slow-growing colonies on cysteine-free medium containing IPTG (after 2 to 3 days, instead of 16 h in the case of its Cys$^+$ parent). Cells in these colonies exhibited the same slow-growth phenotype, which indicated that the mutation was leaky, not highly revertible. The addition of IPTG to DBan41 in cysteine-free liquid medium lengthened the cell doubling time from about 80 min to 280 min.

The cysteine requirement of DBan41 was satisfied by sulfite at 0.3 mM, which indicated a defect in the sulfate assimilation branch of the pathway (Fig. 1). The strain was as sensitive to chromate (MIC, 50 to 100 μM) as its wild-type parent and also grew on low concentrations of thiosulfate (1 to 2 mM). Sulfate uptake mutants are chromate resistant, and many are deficient in thiosulfate uptake (14, 40); therefore this mutation seemed to affect a step leading to sulfite that follows sulfate uptake (Fig. 1).

The mutation was mapped by genetic and molecular methods. (i) The *cys*$^+$ allele was transferred efficiently by Hfr strains BW6164 and CAG5052 to DBan41, which placed the mutation in the 88- to 07-min interval of the *E. coli* chromosome, far from other known cysteine biosynthetic genes (1). (ii) The *cys*$^+$ allele was cotransduced by phage P1 at a frequency of 1% with *zje*::Tn*10*, an insertion at 94 to 95 min (strain BW6458). (iii) The *cys*$^+$ allele was also efficiently transduced by λ656 (23, 24), a λ phage clone that carries the segment of the *E. coli* chromosome from kb ~4511 to kb ~4525 (near 96 min). The mutant allele was recessive to the wild type in partial diploids, which were formed by λ656 infection of a λ$^+$ lysogenic derivative of DBan41 (44), as well as in strains carrying the wild-type allele in multicopy plasmids.

More refined map information came from molecular cloning (Fig. 2). (i) Kan^r plasmids obtained by cloning *Sal*I- or *Cla*I-digested DBan41 DNA in pBR322 contained an 18-kb *Sal*I fragment (pBRGan101) or an overlapping 14-kb *Cla*I fragment (pBRGan102), respectively. (ii) A plasmid obtained by cloning *Sal*I-digested wild-type *E. coli* DNA (pBRGan110) that complemented the *cysQ*::Tn*5tacl* allele contained a 14-kb fragment whose restriction map matched that of the chromosome adjacent to the Tn*5tacl* insertion. (iii) A deletion plasmid that retained only 1.8 kb of chromosomal sequence but retained Cys$^+$ complementing activity was generated by partial *Msp*I digestion of pBRGan110 DNA (pBRGan111). Comparisons of the restriction digest patterns of these clones with the known restriction map of the chromosomal region near 96 min (24, 36) indicated that Tn*5tacl* was at kb 4517, about 900 bp upstream of the *cpdB* gene. Transcription from the *tac* promoter in Tn*5tacl* was toward *cpdB* (clockwise).

Cys$^-$ phenotype not caused by *cpdB* overexpression. The CpdB protein has a 3′-nucleotidase activity that can degrade PAPS to APS in vitro (4). Although CpdB protein seems to be primarily periplasmic, findings of cytoplasmic inhibitors for other periplasmic nucleotidases (36) suggested models in which CpdB also acted intracellularly. Thus, in principle, transcription from the *tac* promoter might cause a cysteine requirement by increasing *cpdB* expression. Alternatively, it might alter the expression of an unknown gene next to *cpdB*. Three findings eliminated the simple CpdB-based model of cysteine auxotrophy: (i) the CpdB level in strain DBan41 was increased less than 2-fold by IPTG (data not shown); (ii) the multicopy *cpdB*$^+$ plasmid (pBRGan104) did not cause a cysteine requirement, although it did result in 10-fold higher CpdB activity (data not shown); and (iii) the cysteine requirement was complemented by plasmid pBRGan102,

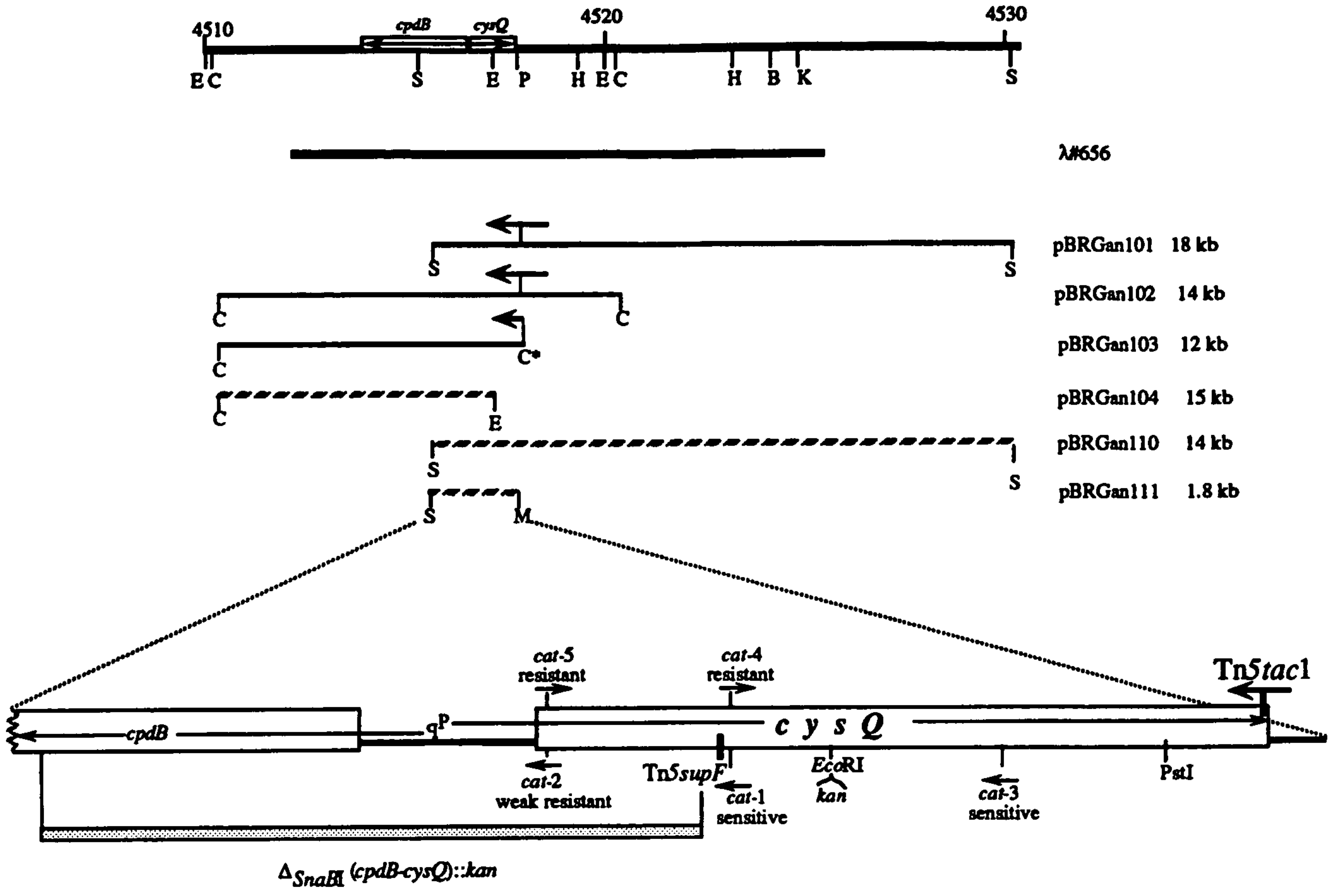


FIG. 2. Physical map of the *cysQ* region of *E. coli*. The extents of bacterial DNA cloned in the phage λ656 and in the plasmids used in this work are depicted, as are the positions of insertions and deletions that helped define *cysQ*. E, *Eco*RI; C, *Cla*I; S, *Sal*I; H, *Hin*dIII, B, *Bam*HI; K, *Kpn*I. The arrows indicate directions of transcription: heavy arrow, from the *tac* promoter in Tn*5tacl*; long arrows, *cysQ* and *cpdB*; smaller arrows, orientations of transcription of a *cat* reporter inserted at *Sau*3A sites in *cysQ*.

which contains a chromosomal segment including *cpdB* and the Tn*5tacl* insertion. These results implied that the cysteine requirement was due to altered expression of a previously unknown gene next to *cpdB*. This gene was designated *cysQ*.

Direction of *cysQ* transcription. Insertions of a *cat* reporter gene were made to determine the orientation of *cysQ* and thereby to deduce whether auxotrophy resulted from overexpression or underexpression of *cysQ* after IPTG-induced transcription from the *tac* promoter in mutant strain DBan41. Five different insertions into *Sau*3A sites of plasmid pBRGan111 that inactivated Cys$^+$ complementation activity were isolated; restriction mapping showed that each was within about 600 bp of the start of *cpdB* (Fig. 2). Insertions 4 and 5, oriented away from *cpdB* (toward Tn*5tacl*), conferred chloramphenicol resistance (25 μg/ml), whereas insertions 1 and 3, in the opposite orientation, did not. Insertion 2, also in the opposite orientation but closest to *cpdB*, conferred weak resistance (~10 μg/ml). This was attributed to a second promoter in *cysQ* that could allow *cpdB* transcription (see sequence analysis, below). Based on insertions 1, 3, 4, and 5, we inferred that *cysQ* is transcribed toward Tn*5tacl*.

Phenotypes conferred by *cysQ* null alleles. Chromosomal *cysQ* null mutations were generated and used to assess whether the distinctive leaky auxotrophy and its correction by anaerobic growth were allele or gene specific. (i) A *cysQ*::Tn*5supF* insertion allele was obtained by selecting transposition of Tn*5supF* to the $cysQ^+$ phage λ656 (29, 43). One of 50 Tn*5supF* insertions resulted in cysteine auxotrophy when recombined into the *E. coli* chromosome, and DNA sequencing (see below) showed that Tn*5supF* was inserted in *cysQ*. (ii) A *cysQ* null allele marked with kanamycin resistance was made by insertion of a *kan* gene at the *Eco*RI site in pBRGan111 (Fig. 2), recombined from the plasmid into λ656, and then recombined from the *cysQ*::*kan* phage into bacterial chromosomes (28). (iii) A segment containing part of both the *cpdB* and the *cysQ* genes was deleted (Δ*Sna*BI; Fig. 2) to further test a possible involvement of *cpdB* in the *cysQ* mutant phenotype. This deletion was marked by insertion of *kan* and recombined into λ656 and from there into bacterial chromosomes. Finally, E. Barnes kindly provided us with a fourth *cysQ* null allele, which was generated by Tn*10* (Tet^r) insertion and was originally designated *amtA*::Tn*10* (21, 22).

Each of these four *cysQ* null alleles resulted in a cysteine requirement that was leaky, corrected by anaerobiosis, and satisfied by sulfite when transduced into several *E. coli* K-12 strain backgrounds, including DB747 (used to isolate the original Tn*5tacl* insertion), DB5508, and MG1655. These alleles were much less leaky in two other laboratory strains: TG1 and ET8000. The strain background determines the leakiness of the *cysQ* mutant phenotype; *cysQ* derivatives of

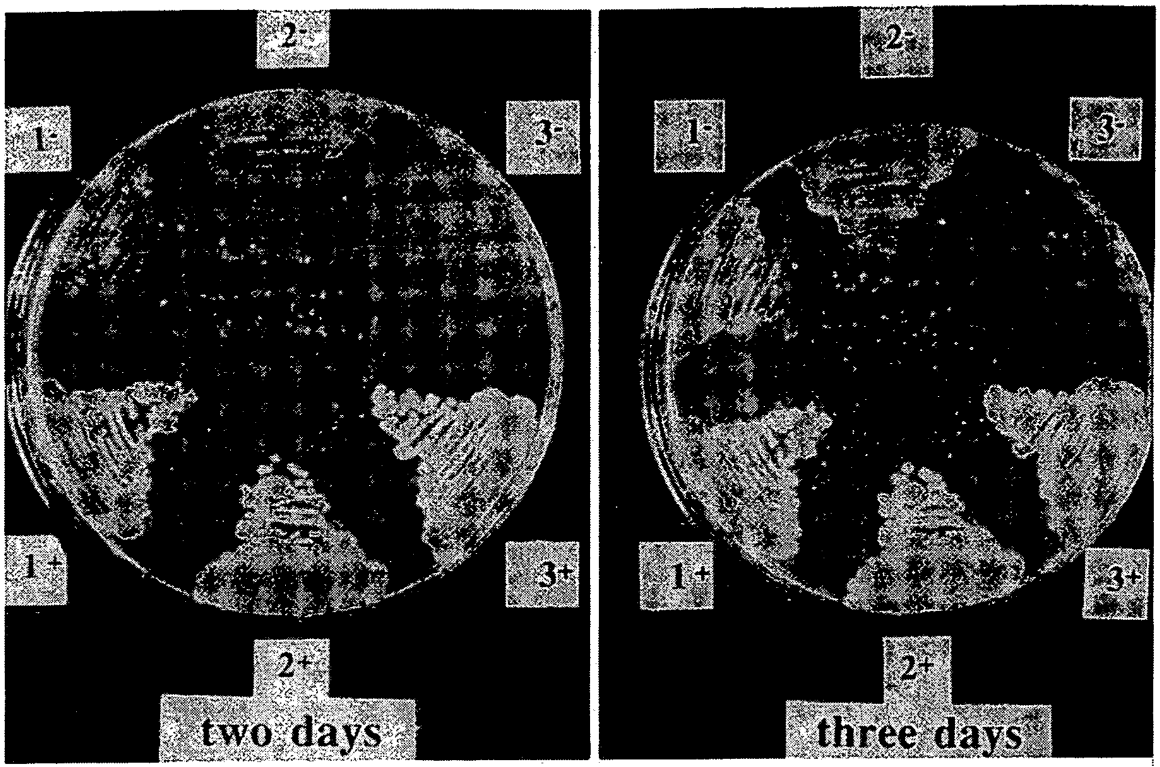


FIG. 3. Growth of *cysQ*::*kan* mutants and wild-type parents. Growth was for 2 or 3 days at 37°C on M9 salts glucose medium (no added cysteine or sulfite). (Scattered white dots near the center of plate are a salt precipitate that often form in M9 solid medium.) The numbers 1, 2, and 3 indicate strain backgrounds DB747, MG1655 and ET8000, respectively. −, *cysQ*::*kan*; +, Cys$^+$.

DB747 and MG1655 showed some growth on cysteine-free medium after 2 days at 37°C and small single colonies after 3 days, whereas the corresponding derivative of ET8000 showed barely perceptible growth only after 3 days (Fig. 3). All *cysQ* mutant strains grew normally on sulfite or under anaerobic conditions. The Cys$^-$ phenotype caused by the Δ(*cysQ cpdB*) allele was identical to that caused by simple insertions in *cysQ* in these strain backgrounds.

The match between the null mutant phenotype and that of the Tn*5tacl* insertion indicated that IPTG-induced transcription from Tn*5tacl* shuts off the expression of *cysQ* quite completely. The equivalence of Cys$^-$ phenotypes of the Δ(*cysQ cpdB*) and the simple *cysQ*::*kan* insertion alleles ruled out a model in which CpdB protein would consume PAPS and in which CysQ protein would regulate this consumption.

DNA sequence of *cysQ*. A 1-kb segment containing the sites of insertion mutations that defined *cysQ* was sequenced by using primer binding sites provided by the Tn*5tacl* and Tn*5supF* transposons, the *kan* and *cat* insertions, and, for one segment, an M13mp18 vector. All portions of this segment were sequenced on both DNA strands. The *cysQ* DNA sequence corresponded to a 246-codon open reading frame preceded by sequences that match consensus transcription promoters and translation initiation sites (Fig. 4). The first part of this sequence matched that found earlier in the 0.5 kb upstream of *cpdB* in *E. coli*; analyses of *S. typhimurium* indicated a *cysQ* homolog in the same location in this species (35). The *cysQ* sequence we determined was identical to that reported for *amtA* (15).

The DNA sequence confirmed that transcription of *cysQ* should diverge from that of *cpdB*, as suggested by *cat* reporter insertions. Tn*5tacl* was inserted within *cysQ*, just two codons from its 3′ end (a fusion protein with 17 additional amino acids is predicted). Tn*5supF* was inserted at codon 72 of *cysQ*. Only 17 bp separate the −35 regions of putative promoters for *cysQ* and *cpdB*, and an apparent consensus cyclic AMP receptor protein (CRP) binding site (placed such that CRP binding could stimulate *cpdB* transcription; see Fig. 4 legend) (35) overlaps the putative *cysQ* promoter. The low-level chloramphenicol resistance associated with the *cat* insertion 2 is attributable to an additional promoter within *cysQ* (nucleotides 72 to 46, underlined in Fig. 4). (This promoter might also explain the relative weakness of the CRP-cyclic AMP dependence of *cpdB* expression [35].) Binding sites for the CysB positive regulatory protein, found in the promoter regions of most cysteine biosynthetic genes (18), did not seem to be present in the *cysQ* promoter region. A search of the PROSITE data base of protein motifs (2) did not reveal significant matches to CysQ. However, we recently found strong amino acid sequence level homologies between *cysQ* and *suhB* of *E. coli* (Fig. 5), mutations in which suppress certain *rpoH* missense alleles (apparently by elevating the levels of heat shock sigma subunit of RNA polymerase that it encodes [56]) and also between *cysQ* and genes for several eukaryotic proteins, including inositol monophosphatase (13, 39).

Does *cysQ* participate in ammonium uptake? While preparing this manuscript, we learned that *cysQ* corresponds to the gene called *amtA* (21, 22), a designation based on the finding of a Tn*10* insertion mutation (*amtA*::Tn*10*) that blocked growth on minimal (cysteine-free) medium containing very low levels of ammonium (≤0.1 mM, rather than the ≥10 mM used in most media). It was proposed that *amtA* is needed

```
                  SD(cpd)
     AATCATCAGGGACATCCTTTTATCATCGGGAATACGAAAGAAAAGGGAGAATAAACGTCT
                      -10(cpd)                      -35(cpd)
     TACTTATAGAACAGTGAAGAATGCCACAATTTTACGCTTTGAAAATGATGACACTATCAC
       -35(cys)                 -10(cys)
     AGTTGGCGCATTCATTAACGATAGGGTATAAGTAAAACAATAAGTTAACACCGCTCACAG
       SD(cys)      1    CAT 2,5
     AGACGAGGTGGAGAAATGTTAGATCAAGTATGCCAGCTTGCACGGAATGCAGGCGATGCC
                    MetLeuAspGlnValCysGlnLeuAlaArgAsnAlaGlyAspAla
      -10                 -35
  46 ATTATGCAGGTCTACGACGGGACGAAACCGATGGACGTCGTCAGCAAAGCGGACAATTCT
     IleMetGlnValTyrAspGlyThrLysProMetAspValValSerLysAlaAspAsnSer

 106 CCGGTAACGGCAGCGGATATTGCCGCTCACACCGTTATCATGGACGGTTTACGTACGCTG
     ProValThrAlaAlaAspIleAlaAlaHisThrValIleMetAspGlyLeuArgThrLeu
                                CAT 1,4    (I)Tn5supF(O)
 166 ACACCGGATGTTCCGGTCCTTTCTGAAGAAGATCCTCCCGGTTGGGAAGTCCGTCAGCAC
     ThrProAspValProValLeuSerGluGluAspProProGlyTrpGluValArgGlnHis

 226 TGGCAGCGTTACTGGCTGGTAGACCCGCTGGATGGTACTAAAGAGTTTATTAAACGTAAT
     TrpGlnArgTyrTrpLeuValAspProLeuAspGlyThrLysGluPheIleLysArgAsn
       EcoRI
 286 GGCGAATTCACCGTTAACATTGCGCTCATTGACCATGGCAAACCGATTTTAGGCGTGGTG
     GlyGluPheThrValAsnIleAlaLeuIleAspHisGlyLysProIleLeuGlyValVal

 346 TATGCGCCGGTAATGAACGTAATGTACAGCGCGGCAGAAGGCAAAGCGTGGAAAGAAGAG
     TyrAlaProValMetAsnValMetTyrSerAlaAlaGluGlyLysAlaTrpLysGluGlu
                                                          CAT3
 406 TGCGGTGTGCGCAAGCAGATTCAGGTCCGCGATGCGCGCCCGCCGCTGGTGGTGATCAGC
     CysGlyValArgLysGluIleGlnValArgAspAlaArgProProLeuValValIleSer

 466 CGTTCCCATGCGGATGCGGAGCTGAAAGAGTATCTGCAACAGCTTGGCGAACATCAGACC
     ArgSerHisAlaAspAlaGluLeuLysGluTyrLeuGlnGlnLeuGlyGluHisGlnThr

 526 ACGTCCATCGGCTCTTCGCTGAAATTCTGCCTGGTGGCGGAAGGACAGGCGCACGTGTAC
     ThrSerIleGlySerSerLeuLysPheCysLeuValAlaGluGlyGlnAlaHisValTyr
                                                          PstI
 586 CCGCGCTTCGGACCAACGAATATTTGGGACACCGCCGCTGGACATGCTGTAGCTGCAGCT
     ProArgPheGlyProThrAsnIleTrpAspThrAlaAlaGlyHisAlaValAlaAlaAla

 646 GCCGGAGCGCACGTTCACGACTGGCAGGGTAAACCGCTGGATTACACTCCGCGTGAGTCG
     AlaGlyAlaHisValHisAspTrpGlnGlyLysProLeuAspTyrThrProArgGluSer
                       (O)Tn5tac1(I)    741
 706 TTCCTGAATCCGGGGTTCAGAGTGTCTATTTACTAAATTCAGATGGCAGAAACAGTGTAT
     PheLeuAsnProGlyPheArgValSerIleTyrEnd

     TTCCTGATTCTGCCATCCTGATTTCTCCCAACCTAAAAAGTTATAAATAAAAAGAGATTG

     TATTTAAAGTGCAAAAATTCAATTGCTAATAAGTTACA
```

FIG. 4. DNA sequence of the *cysQ* gene. Sites of insertion and orientations (in cases of transposons) are indicated, as are putative promoters for *cysQ* and *cpdB* and the consensus CRP binding site. The CRP binding site identified in ref. 35 extends from positions −72 to −96, relative to the start of *cysQ* translation (beginning at TT in the −35 region of the *cysQ* promoter). The DNA sequencing protocols and sequences of the oligonucleotide primers used are given in Materials and Methods.

for active ammonium uptake (21). Our reconstruction experiments showed, however, that even *cysQ*$^+$ (*amtA*$^+$) bacteria grew very poorly on low-ammonium medium (Fig. 6). Strains with the *amtA*::Tn*10* or *cysQ*::*kan* insertion mutations failed to form colonies on this medium, as reported earlier (21). The mutants did grow, however, when this medium was supplemented with cysteine (Fig. 6) or sulfite, in which case these mutant strains were indistinguishable from their Cys$^+$ parents in colony size. Since neither sulfite nor cysteine can be used as an ammonium source by *E. coli* (53), these results are not consistent with the interpretation (21) that the AmtA (CysQ) protein is needed for ammonium uptake.

Possible roles for *cysQ*. We tested whether *cysQ* might be needed in sulfate uptake or activation. No effect of *cysQ* null mutations was found on the rapid uptake of labelled sulfate from the medium in the background of strain TG1 or DB5463 (nonleaky and leaky cysteine requirements, respectively). In contrast, a *cysDN* (ATP sulfurylase) deletion strain was severely deficient in sulfate uptake, as expected (14) (data not shown). No significant differences between *cysQ* mutant and parental strains were detected in levels of ATP sulfurylase, which catalyzes synthesis of APS, the first activated sulfur intermediate. *cysQ* mutations also had no effect on the level of the activator of ATP sulfurylase (data not shown).

Reversion of *cysQ* null mutants. Cys$^+$ revertants of

```
   MLDQVCQLARNAGDAIMQVYDGTKPMDVVSKADNSPVTAADIAAHTVIMDG      (1-51)    CysQ
   ML      AR AG  I   Y+    ++   K+ N  VT  D AA   VI+D
MHPMLNIAVRAARKAGNLIAKNYETPDAVEASQKGSNDFVTNVDKAAEAVIIDT      (1-54)    SuhB

LRTLTPDVPVLSEEDPPGWEVRQHWQRYWLVDPLDGTKEFIKRNGEFTVNIALI      (52-105)  CysQ
+R   P   +++EE     E        W++DPLDGT  FIKR   F V IA+
IRKSYPQHTIITEE-SGELEGTDQ-DVQWVIDPLDGTTNFIKRLPHFAVSIAVR      (55-106)  SuhB

DHGKPILGVVYAPVMNVMYSAAEGKAWKEECGVRK-QIQVRDARPPLVVISRSH      (106-157) CysQ
 +G+  ++VVY P+ N +++A  G +     G R      RD    ++
IKGRTEVAVVYDPMRNELFTATRGQG-AQLNGYRLLGSTARDLDGTILATGFPF      (107-159) SuhB

ADAEL KEYLQQLGE-----HQTTSIGSS-LKFCLVAEGQAHVYPRFGPTNIWD      (158-205) CysQ
        Y++ +G             GS  L    VA G    +   G    WD
KAKQYATTYINIVGKLFNECADFRRTGSAALDLAYVAAGRVDGFFEIG-LRPWD      (160-212) SuhB

TAAGHAVAAAAGAHVHDWQGKPLDYTPRESFLNPGFRVSIY                   (206-246) CysQ
 AAG  +   AG+ V D  G    Y      +    RV
FAAGELLVREAGGIVSDFTGGH-NYMLTGNIVAGNPRVVKAMLANMRDELSDALKR    (213-267) SuhB
```

FIG. 5. Alignment of amino acid sequences of inferred protein products of *cysQ* and *suhB* (adapted from data in reference 39). Identities are indicated by placements of conserved amino acids in the middle line; conservative substitutions (+) are indicated. Overlined segments indicate regions with high sequence similarity to inositol monophosphatase (13). The CysQ (SuhB) initial amino acid alignment score of 229, calculated by using the FASTA program (42), was 28 standard deviation units above the mean initial score of 24.6 for comparisons of CysQ to the other sequences in the PIR (release 28) protein data base. In a test using the Dayhoff Relate program, there were four segments of 25 amino acids in length that were more than eight standard deviations above the mean. This indicates strong homology: the probability of getting a single segment with such a deviation from a random sequence by chance alone is less than 10^{-15} (42).

cysQ::*kan* and *cysQ*::Tn*10* mutants were obtained at frequencies of about 10^{-6} with derivatives of TG1 and ET8000 (nonleaky *cysQ* mutant phenotype) and $>10^{-5}$ with derivatives of DB747 and MG1655 (leaky *cysQ* mutant phenotype); these differences in recovery probably reflect the greater leakiness of *cysQ* mutations in the DB747 and MG1655 backgrounds. The revertants were heterogeneous in colony size on cysteine-free medium but grew as well as their $cysQ^+$ ancestors on cysteine-containing medium and retained the Kan^r or Tet^r traits of their Cys$^-$ parents. The parental *cysQ* mutant alleles were recovered from several revertants by transduction and selection for the appropriate resistance trait (Kan^r or Tet^r). Two spontaneous reversion mutations that allowed relatively good growth on cysteine-free medium were mapped in Hfr × F$^-$ crosses and then by transduction with several candidate λ phage clones (marked by insertion of a Tn*5cam* transposon [44, 50]). These reversion mutations were found to be in the segment carried by λ419, a phage clone that also carries the *cysPTWA* (sulfate binding and uptake) operon. The two revertants tested were found to be defective in sulfate uptake, unlike their cysteine-requiring parents. Partial diploids generated by lysogenizing revertants with λ419::Tn*5cam* and a λ$^+$ helper required cysteine, indicating that the reversion mutation is recessive and thus probably due to loss of function. This Cys$^-$ phenotype was unstable, however, because of frequent homogenotization for the parental (nonrevertant) allele.

DISCUSSION

The initial steps in the sulfate assimilation branch of the cysteine pathway entail sulfate uptake, its activation via formation of APS and conversion to PAPS, and then its reduction to sulfite (Fig. 1). The mutational and sequence analyses presented here identified a previously uncharacterized gene, *cysQ*, whose product is needed for proper metabolic functioning of this part of the pathway. We propose below that CysQ acts on PAPS. The *cysQ* gene was mapped to a locus at ~96 min in the *E. coli* chromosome, which is far from other *cys* genes. The *cysQ* promoter region overlapped a CAP binding site that is implicated in the control of expression of the adjacent gene, *cpdB* (35), and it did not contain a good match to the consensus binding site for the CysB regulatory protein (18). Hence, the expression of *cysQ* may be controlled differently from that of most other *cys* genes. The auxotrophy resulting from mutations in *cysQ* was leaky in some strain backgrounds and was compensated by mutations in other genes; *cysQ* mutants were prototrophic during anaerobic growth.

A precedent for cysteine biosynthetic genes that are not needed during anaerobic growth is provided by *cysI* and *cysJ* in *S. typhimurium* (3). This case reflects the presence of additional anaerobic sulfite reduction (*asr*) genes. *E. coli* lacks such *asr* genes (20), however, and our studies indicate that *cysQ* acts before, not after, sulfite formation (Fig. 1). *cysB* is also only needed during aerobic growth (3), suggesting that a separate transcriptional activator may be operating anaerobically. *cysQ* does not seem to be a transcriptional activator of *cys* genes, since it does not significantly affect the rate of sulfate uptake or the level of ATP sulfurylase or its protein activator.

cysQ is identical to *amtA*, a gene which had been thought to participate in ammonium transport (21). That interpretation was based on a failure of mutant strains to grow on low-ammonium cysteine-free medium or to take up methylammonium at a high rate when they were grown with arginine in place of ammonium (21). We found that the growth defect of *cysQ* (*amtA*) mutants was compensated by sulfite or cysteine (Fig. 6), neither of which serves as a nitrogen source (53). The initial failure to recognize the cysteine requirement of the *amtA* mutant (21) may have been due to leakiness on normal minimal medium (Fig. 3) or inadvertent selection of a (partially) compensating suppressor mutation. The inability of *cysQ* (*amtA*) mutants to grow on low-ammonium medium probably results from the combined effects of partial starvation for ammonium (because of

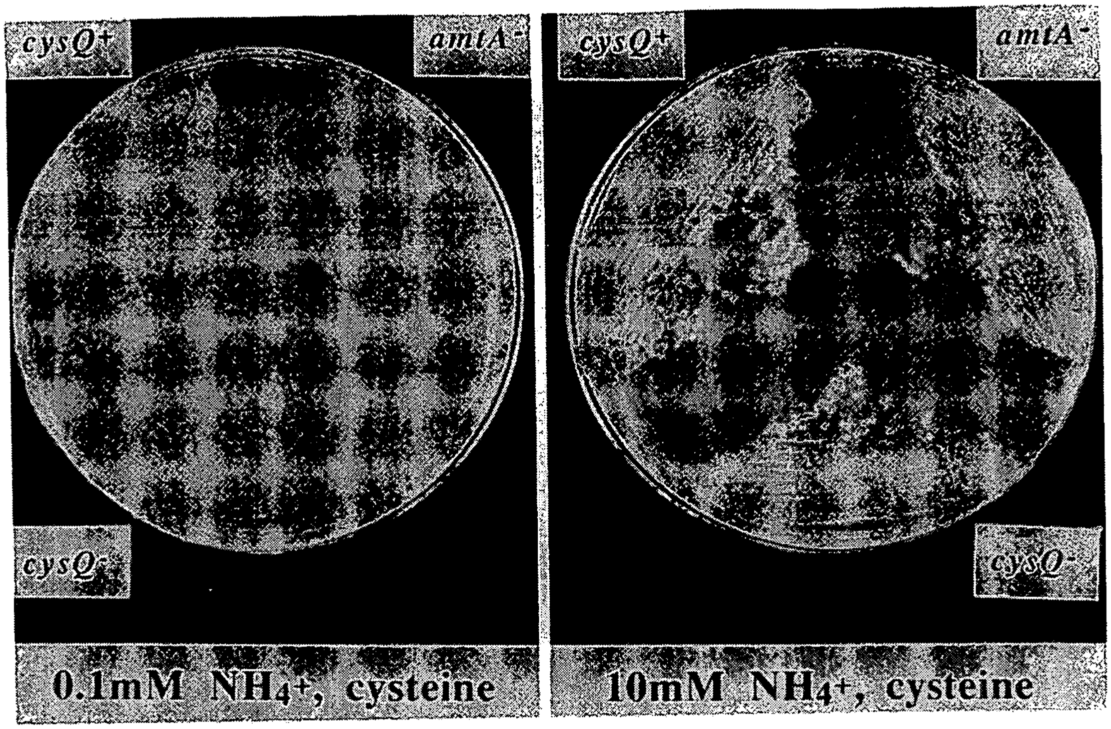


FIG. 6. Comparison of wild-type and *cysQ* and *amtA* mutant bacterial strains on modified M9 minimal medium and either low (0.1 mM) or normal (10 mM) concentrations of ammonium acetate. Strains: *cysQ*$^+$, ET8000; *cysQ* mutant, ET8000 *cysQ*::*kan* (DB6908); *amtA* mutant, ET8000 *amtA*::Tn*10* (AJ2653) grown in medium containing 0.2 mM cysteine. The plate with 10 mM ammonium was incubated for 1 day and the plate with 0.1 mM ammonium was incubated for 2 days at 37°C before being photographed. Identical growth patterns were obtained with sulfite in place of cysteine, but only Cys$^+$ parental strains grew on minimal low-ammonium medium lacking sulfite or cysteine. Equivalent weak growth of mutant and Cys$^+$ sibling strains on the low-ammonium medium was also observed with all other lineages tested (DB747 and its *cysQ* derivative DB5656, MG1655 and its *cysQ* derivative DB6316, TG1 and its *cysQ* derivative DB6913). Growth was weaker on medium containing 0.025 mM rather than 0.1 mM ammonium acetate, as expected, since ammonium was limiting. In the cases of Cys$^+$ strains, this slow growth was not stimulated by adding 10 or 20 mM cysteine (or sulfite or thiosulfate), whereas growth was stimulated by adding 20 mM glutamate or arginine (which serve as ammonium sources [53]).

the medium) and for cysteine (because of a mutation). Although the inefficient induced methylammonium uptake by *amtA* cells was also interpreted to reflect a specific uptake defect, the reported data (21) indicate that the basal level of uptake was not affected by the *amtA* mutation. Earlier work had shown that induction by growth in arginine reflects the slow release of ammonium from this source, relative to the rate of ammonium consumption (45, 53). Because the poor growth of *cysQ* (*amtA*) mutants on cysteine-free medium should allow the arginine-derived ammonium to accumulate to repressing levels, we do not find it necessary to postulate a role for CysQ (AmtA) in ammonium or methylammonium uptake.

How does CysQ act in the synthesis of sulfite and cysteine? Several possible roles have been eliminated by our results to date: (i) sulfate uptake, (ii) stabilization of ATP sulfurylase, (iii) synthesis or stabilization of the ATP sulfurylase activator, and (iv) modulation of CpdB. In addition, the *cysQ* sequence does not match that of *ppa*, the gene for pyrophosphatase (30), an enzyme probably needed for efficient APS synthesis (Fig. 1). The leakiness of *cysQ*-null alleles in many strain backgrounds might reflect (i) a second gene with a functionally related role; (ii) an intrinsic activity of gene(s) that can mutate to give a Cys$^+$ revertant phenotype; or, if *cysQ* is regulatory, (iii) strain background-dependent differences in the quantitative effects of CysQ on the gene, protein, or metabolite that is the target of its control.

Studies of Cys$^+$ revertants are providing insights into how CysQ may act. Several spontaneous reversion mutations were mapped to a region that includes *cysTWA* permease genes, were recessive to the wild-type (nonrevertant) alleles, and were defective in sulfate uptake. These results suggested that reversion results from loss of function, not from an unusual expression of a silent or cryptic suppressor gene. Accordingly, we have begun to isolate transposon insertions that restore prototrophy to *cysQ* mutants in a nonleaky background (50). One insertion that resulted in very small colonies on cysteine-free medium was in *cysA*, which encodes a subunit of sulfate permease (49). Transduction of this insertion into a *cysQ*$^+$ strain resulted in the same small-colony phenotype, indicating that the phenotype reflected loss of *cysA* function, not poor suppression of the *cysQ* mutation. A second insertion, which resulted in colonies of nearly normal size, was in *cysP*, a gene whose product contributes to efficient sulfate and thiosulfate binding (19). In interpreting these reversion data we draw on early findings that mutations in *cysH* or in *trxA* plus *grx* cause poor growth, apparently because accumulated PAPS or a derivative of it is toxic, and that these mutations can be compensated by mutations inactivating sulfate permease (16, 45a). Although the role of *cysP* in the *cys* pathway is not understood, the ability of permease mutations to compensate for the defect in *cysQ* suggests that CysQ also acts on PAPS. Perhaps CysQ participates with the CysH sulfotransferase to generate sulfite. Alternatively, perhaps CysQ sequesters or consumes excess PAPS or a toxic derivative of it. On this latter view, cysteine might be needed for growth of *cysQ* mutants only to allow repression of *cys* gene expression and thereby decrease PAPS synthesis, rather than to compensate for a missing biosynthetic enzyme. CysQ exhibits striking amino acid sequence homology to mammalian inositol monophosphatase as well as to the product of the *suhB* gene of *E. coli* (Fig. 6) (39). The homology between CysQ and inositol monophosphatase, in particular, encourages models in which CysQ acts on a phosphorylated metabolite such as PAPS, possibly ensuring that it plays its essential biosynthetic role without toxicity to the cell.

ACKNOWLEDGMENTS

We thank C. M. Berg, H. V. Huang, S. H. Phadnis, M. Russel, and P. M. Wikström for stimulating discussions and A. Jayakumar and E. Barnes for the exchange of strains and unpublished information.

A.F.N. was a trainee in Medicine (Infectious Diseases) (Public Health Service) training grant 5 T32 AI07172 from the National Institute of Allergy and Infectious Diseases) and in Medical Informatics (NLM training grant 5-TI-LM07049). This work was supported by Public Health Service grants GM-37138, HG00563, BRSG 613-4743, and DK43003 from the National Institutes of Health and by grant DEFGO2-89ER60862 from the U.S. Department of Energy to D.E.B. and T.S.L.

REFERENCES

1. **Bachmann, B. J.** 1990. Linkage map of *Escherichia coli* K-12, edition 8. Microbiol. Rev. **54:**130–197.
2. **Bairoch, A.** 1991. PROSITE: a dictionary of sites and patterns in proteins. Nucleic Acids Res. **19:**2241–2245.
3. **Barret, E. L., and G. W. Chang.** 1979. Cysteine auxotrophs of *Salmonella typhimurium* which grow without cysteine in a hydrogen/carbon dioxide atmosphere. J. Gen. Microbiol. **115:** 513–516.
4. **Beacham, I. R., and S. Garrett.** 1980. Isolation of *Escherichia coli* mutants (*cpdB*) deficient in periplasmic 2′:3′-cyclic phosphodiesterase and genetic mapping of the *cpdB* locus. J. Gen. Microbiol. **119:**31–34.
5. **Berg, D. E., A. Weiss, and L. Crossland.** 1980. Polarity of Tn*5* insertion mutations in *Escherichia coli*. J. Bacteriol. **142:**439–446.
6. **Bolivar, F., R. L. Rodriguez, P. J. Greene, M. C. Betlach, H. L. Heyneker, and H. W. Boyer.** 1977. Construction and characterization of new cloning vehicles. II. A multipurpose cloning system. Gene **2:**95–113.
7. **Campbell, A.** 1961. Sensitive mutants of bacteriophage λ. Virology **14:**22–32.
8. **Casadaban, M. J., and S. N. Cohen.** 1980. Analysis of gene control signals by DNA fusion and cloning in *Escherichia coli*. J. Mol. Biol. **138:**179–207.
9. **Chow, W.-Y., and D. E. Berg.** 1988. Tn*5tac1*, a derivative of transposon Tn*5* that generates conditional mutations. Proc. Natl. Acad. Sci. USA **85:**6468–6472.
10. **Close, T. J., and R. L. Rodriguez.** 1982. Construction and characterization of the chloramphenicol-resistance gene cartridge: a new approach to the transcriptional mapping of extrachromosomal elements. Gene **20:**305–316.
11. **Csonka, L. N., and A. J. Clark.** 1979. Deletions generated by the transposon Tn*10* in the *srl recA* region of the *Escherichia coli* K-12 chromosome. Genetics **93:**321–343.
12. **Davis, R. W., D. Botstein, and J. Roth.** 1980. Advanced bacterial genetics. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
13. **Diehl, R. E., P. Whiting, J. Potter, N. Gee, C. I. Ragan, D. Linemeyer, R. Schoepfer, C. Bennett, and R. A. F. Dixon.** 1990. Cloning and expression of bovine brain inositol monophosphatase. J. Biol. Chem. **265:**5946–5949.
14. **Dreyfuss, J.** 1964. Characterization of a sulfate and thiosulfate-transporting system from *Salmonella typhimurium*. J. Biol. Chem. **239:**2292–2297.
15. **Fabiny, J. M., A. Jayakumar, A. C. Chinault, and E. M. Barnes.** 1991. Ammonium transport in *Escherichia coli*: localization and nucleotide sequence of the *amtA* gene. J. Gen. Microbiol. **137:**983–989.
16. **Gillespie, D., M. Demerec, and H. Itakawa.** 1968. Appearance of double mutants in aged cultures of *Salmonella typhimurium* cysteine-requiring strains. Genetics **59:**433–442.
17. **Guyer, M. S., R. R. Reed, J. A. Steitz, and K. B. Low.** 1981. Identification of a sex-factor-affinity site in *E. coli* as $\gamma\delta$. Cold Spring Harbor Symp. Quant. Biol. **45:**135–139.
18. **Hrniewicz, M., and N. M. Kredich.** 1991. The *cysP* promoter of *Salmonella typhimurium*: characterization of two binding sites for CysB protein, studies of in vivo transcription initiation, and demonstration of anti-inducer effects of thiosulfate. J. Bacteriol. **173:**5876–5886.
19. **Hrniewicz, M., A. Sirko, A. Palucha, A. Bock, and D. Hulanicka.** 1990. Sulfate and thiosulfate transport in *Escherichia coli* K-12: identification of a gene encoding a novel protein involved in thiosulfate binding. J. Bacteriol. **172:**3358–3366.
20. **Huang, C. J., and E. L. Barrett.** 1990. Identification and cloning of genes involved in anaerobic sulfite reduction by *Salmonella typhimurium*. J. Bacteriol. **172:**4100–4102.
21. **Jayakumar, A., S. J. Hwang, J. M. Fabiny, A. C. Chinault, and E. M. Barnes.** 1989. Isolation of an ammonium or methylammonium ion transport mutant of *Escherichia coli* and complementation by the cloned gene. J. Bacteriol. **171:**996–1001.
22. **Jayakumar, A., K. E. Rudd, J. M. Fabiny, and E. M. Barnes.** 1991. Localization of the *Escherichia coli amtA* gene to 95.8 minutes. J. Bacteriol. **173:**1572–1573.
23. **Kohara, Y.** 1990. Correlation between the physical and genetic maps of the *Escherichia coli* K-12 chromosome, p. 29–42. *In* K. Drlica and M. Riley (ed.), The bacterial chromosome. American Society for Microbiology, Washington, D.C.
24. **Kohara, Y., K. Akiyama, and K. Isono.** 1987. The physical map of the whole *E. coli* chromosome: application of a new strategy for rapid analysis and sorting of a large genomic library. Cell **50:**495–508.
25. **Kredich, N. M.** 1987. Biosynthesis of cysteine, p. 419–428. *In* F. C. Neidhardt, J. L. Ingraham, K. B. Low, B. Magasanik, M. Schaechter, and H. E. Umbarger (ed.), *Escherichia coli* and *Salmonella typhimurium*: cellular and molecular biology. American Society for Microbiology, Washington, D.C.
26. **Kredich, N. M., L. J. Foote, and M. D. Hulanicka.** 1975. Studies on the mechanism of inhibition of *Salmonella typhimurium* by 1,2,4-triazole. J. Biol. Chem. **250:**7324–7331.
27. **Krishnan, B. R., D. Kersulyte, I. Brikun, H. V. Huang, C. M. Berg, and D. E. Berg.** 1992. Transposon- and PCR-based sequencing of DNAs cloned in λ phage. Methods Enzymol., in press.
28. **Kulakauskas, S., P. M. Wikström, and D. E. Berg.** 1991. Efficient introduction of cloned mutant alleles into the *Escherichia coli* chromosome. J. Bacteriol. **173:**2633–2638.
29. **Kurnit, D. M., and B. Seed.** 1990. Improved genetic selection for screening bacteriophage libraries by homologous recombination in vivo. Proc. Natl. Acad. Sci. USA **87:**3166–3169.
30. **Lahti, R., T. Pitkäranta, E. Valve, I. Ilta, E. Kukko-Kalske, and J. Heinonen.** 1988. Cloning and characterization of the gene encoding inorganic pyrophosphatase of *Escherichia coli* K-12. J. Bacteriol. **170:**5901–5907.
31. **Leyh, T. S.** 1990. The discovery of a coenzyme that stimulates the activity of ATP sulfurylase from *E. coli* K-12. FASEB J. **4:**A1988 (abstr. 1717).
32. **Leyh, T. S., and G. D. Markham.** Unpublished data.
33. **Leyh, T. S., and Y. Suo.** 1992. GTPase mediated activation of ATP sulfurylase. J. Biol. Chem., in press.
34. **Leyh, T. S., J. C. Taylor, and G. D. Markham.** 1988. The sulfate activation locus of *Escherichia coli* K-12: cloning, genetic, and enzymatic characterization. J. Biol. Chem. **263:**2409–2416.
35. **Liu, J., and I. R. Beacham.** 1990. Transcription and regulation of the *cpdB* gene in *Escherichia coli* K12 and *Salmonella typhmurium* LT2: evidence for modulation of constitutive promoters by

cyclic AMP-CRP complex. Mol. Gen. Genet. 222:161–165.

36. **Liu, J., D. M. Burns, and I. R. Beacham.** 1986. Isolation and sequence analysis of the gene (*cpdB*) encoding periplasmic 2′,3′-cyclic phosphodiesterase. J. Bacteriol. **165:**1002–1010.
37. **MacNeil, T., D. MacNeil, and B. Tyler.** 1982. Fine-structure map and complementation analysis of the *glnA-glnL-glnG* region in *Escherichia coli*. J. Bacteriol. **150:**1302–1313.
38. **Messing, J.** 1983. New M13 vectors for cloning. Methods Enzymol. **101:**20–78.
39. **Neuwald, A. F., J. D. York, and P. W. Majerus.** 1991. Diverse proteins homologous to inositol monophosphatase. FEBS Lett. **294:**16–18.
40. **Pardee, A. B., L. S. Prestidge, M. B. Whipple, and J. Dreyfuss.** 1966. A binding site for sulfate and its relation sulfate transport into *Salmonella typhimurium*. J. Biol. Chem. **241:**3962–3969.
41. **Parra, F., P. Britton, C. Castle, M. C. Jones-Mortimer, and H. L. Kornberg.** 1983. Two separate genes involved in sulphate transport in *Escherichia coli* K12. J. Gen. Microbiol. **129:**357–358.
42. **Pearson, W. R., and D. J. Lipman.** 1988. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA **85:**2444–2448.
43. **Phadnis, S. H., H. V. Huang, and D. E. Berg.** 1989. Tn*5supF*, a 264-base-pair transposon derived from Tn*5* for insertion mutagenesis and sequencing DNAs cloned in phage λ. Proc. Natl. Acad. Sci. USA **86:**5908–5912.
44. **Phadnis, S. H., S. Kulakauskas, B. R. Krishnan, J. Hiemstra, and D. E. Berg.** 1991. Transposon Tn*5supF*-based reverse genetic method for mutational analysis of *Escherichia coli* with DNAs cloned in λ phage. J. Bacteriol. 173:896–899.
45. **Reitzer, L. J., and B. Magasanik.** 1989. Ammonia assimilation and the biosynthesis of glutamine, glutamate, aspartate, asparagine, L-alanine, and D-alanine, p. 302–320. *In* F. C. Neidhardt J. L. Ingraham, K. B. Low, B. Magasanik, M. Schaechter, and H. E. Umbarger (ed.), *Escherichia coli* and *Salmonella typhimurium*: cellular and molecular biology. American Society for Microbiology, Washington, D.C.

45a. **Russel, M., P. Model, and A. Holmgren.** 1990. Thioredoxin or glutaredoxin in *Escherichia coli* is essential for sulfate reduction, but not for deoxyribonucleotide synthesis. J. Bacteriol. **172:**1923–1929.

46. **Satishchandran, C., and G. D. Markham.** 1989. Adenosine-5′-phosphosulfate kinase from *Escherichia coli* K-12. J. Biol. Chem. **264:**15012–15021.
47. **Sambrook, J., E. F. Fritsch, and T. Maniatis.** 1989. Molecular cloning: a laboratory manual, 2nd ed. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
48. **Singer, M., T. A. Baker, G. Schnitzler, S. M. Deischel, M. Goel, W. Dove, K. J. Jaacks, A. D. Grossman, J. W. Erickson, and C. A. Gross.** 1989. A collection of strains containing genetically linked alternating antibiotic resistance elements for genetic mapping of *Escherichia coli*. Microbiol. Rev. **53:**1–24.
49. **Sirko, A., M. Hrniewicz, D. Hulanicka, and A. Bock.** 1990. Sulfate and thiosulfate transport in *Escherichia coli* K-12: nucleotide sequence and expression of the *cysTWAM* gene cluster. J. Bacteriol. **172:**3351–3357.
50. **Suziedelis, K., S. Kulakauskas, and D. E. Berg.** Unpublished data.
51. **Tsang, M. L.-S., and J. A. Schiff.** 1976. Sulfate-reducing pathway in *Escherichia coli* involving bound intermediates. J. Bacteriol. **125:**923–933.
52. **Tsang, M. L.-S.** 1981. Assimilatory sulfate reduction in *Escherichia coli*: identification of the alternate cofactor for adenosine 3′-phosphate 5′-phosphosulfate reductase as glutaredoxin. J. Bacteriol. **146:**1059–1066.
53. **Tyler, B.** 1978. Regulation of the assimilation of nitrogen compounds. Annu. Rev. Biochem. **47:**1127–1162.
54. **Vogel, H. J., and D. M. Bonner.** 1956. Acetylornithinase of *Escherichia coli*: partial purification and some properties. J. Biol. Chem. **218:**97–106.
55. **Wanner, B. L.** 1986. Novel regulatory mutants of the phosphate regulon of *Escherichia coli* K-12. J. Mol. Biol. **191:**39–58.
56. **Yano, R., H. Nagain, K. Shiba, and T. Yura.** 1990. A mutation that enhances synthesis of σ^{32} and suppresses temperature-sensitive growth of the *rpoH15* mutant of *Escherichia coli*. J. Bacteriol. 172:2124–2130.

APPENDIX K

APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Apr. 2003, p. 2006–2014 Vol. 69, No. 4
0099-2240/03/$08.00+0 DOI: 10.1128/AEM.69.4.2006–2014.2003

Identification of a Third Sulfate Activation System in *Sinorhizobium* sp. Strain BR816: the CysDN Sulfate Activation Complex

Carla Snoeck,[1] Christel Verreth,[1] Ismael Hernández-Lucas,[2] Esperanza Martínez-Romero,[2] and Jos Vanderleyden[1]*

Centre of Microbial and Plant Genetics, Heverlee, Belgium,[1] *and Centro de Investigación sobre Fijación de Nitrógeno, Cuernavaca, Morelos, Mexico*[2]

Received 30 September 2002/Accepted 15 January 2003

***Sinorhizobium* sp. strain BR816 possesses two *nodPQ* copies, providing activated sulfate (3′-phosphoadenosine-5′-phosphosulfate [PAPS]) needed for the biosynthesis of sulfated Nod factors. It was previously shown that the Nod factors synthesized by a *nodPQ* double mutant are not structurally different from those of the wild-type strain. In this study, we describe the characterization of a third sulfate activation locus. Two open reading frames were fully characterized and displayed the highest similarity with the *Sinorhizobium meliloti* housekeeping ATP sulfurylase subunits, encoded by the *cysDN* genes. The growth characteristics as well as the levels of Nod factor sulfation of a *cysD* mutant (FAJ1600) and a *nodP1 nodQ2 cysD* triple mutant (FAJ1604) were determined. FAJ1600 shows a prolonged lag phase only with inorganic sulfate as the sole sulfur source, compared to the wild-type parent. On the other hand, FAJ1604 requires cysteine for growth and produces sulfate-free Nod factors. Apigenin-induced *nod* gene expression for Nod factor synthesis does not influence the growth characteristics of any of the strains studied in the presence of different sulfur sources. In this way, it could be demonstrated that the "household" CysDN sulfate activation complex of *Sinorhizobium* sp. strain BR816 can additionally ensure Nod factor sulfation, whereas the symbiotic PAPS pool, generated by the *nodPQ* sulfate activation loci, can be engaged for sulfation of amino acids. Finally, our results show that rhizobial growth defects are likely the reason for a decreased nitrogen fixation capacity of bean plants inoculated with *cysD* mutant strains, which can be restored by adding methionine to the plant nutrient solution.**

Sulfur is a macronutrient that is required by all organisms. It forms constituents of proteins, lipids, carbohydrates, electron carriers, and numerous cellular metabolites. Sulfate is the most abundant source of utilizable sulfur in the aerobic biosphere. The sulfate assimilation complex, required for the formation of the sulfur-containing amino acid cysteine, has been the subject of intensive study in *Escherichia coli* (21). Cysteine is the central precursor of all organic molecules containing reduced sulfur, ranging from the amino acid methionine to peptides, proteins, vitamins, cofactors such as *S*-adenosylmethionine, and hormones.

Like all inorganic nutrients, sulfate is transported into cells by highly specific membrane transport systems (18). Sulfate assimilation requires its prior activation to adenylate compounds via a pathway that seems to be similar in all organisms. The activation is achieved by the ATP sulfurylase-catalyzed reaction of sulfate with ATP to give adenosine 5′-phosphosulfate (APS), coupled with GTP hydrolysis. Subsequently, APS is phosphorylated by an APS kinase to produce 3′-phosphoadenosine-5′-phosphosulfate (PAPS). In *E. coli*, ATP sulfurylase is encoded by *cysD* and *cysN*, whereas the APS kinase is encoded by *cysC* (27, 28). PAPS is then enzymatically reduced by the *cysH*-encoded PAPS reductase (also known as PAPS sulfotransferase) to sulfite, which enters the cysteine biosynthetic pathway.

PAPS also serves directly as a sulfate donor for the formation of sulfated compounds. For example, *Rhizobium*-legume symbiotic interactions are mediated by a host-specific bacterial signaling molecule (the Nod factor), which can be sulfated. In general, rhizobial species that produce sulfated Nod factors possess at least two sulfate activation systems (6, 12, 24, 25, 40). The three genes that are indispensable for Nod factor sulfation, *nodP*, *nodQ*, and *nodH*, were first isolated from *Sinorhizobium meliloti*. Together, *nodP* and *nodQ* encode both ATP sulfurylase and APS kinase activities (45, 47), whereas the *nodH* gene product, a sulfotransferase, directly transfers the activated sulfate moiety to the Nod factor backbone (8, 44). NodP is homologous to *E. coli* CysD, while the amino- and carboxy-terminal domains of NodQ are homologous to *E. coli* CysN and CysC, respectively. In a recent study, it was reported that the specificity of phytopathogen-host interactions also can be controlled by a sulfated avirulence effector molecule, which is yet to be identified (48). The rice pathogen *Xanthomonas oryzae* pv. oryzae RaxP and RaxQ proteins are responsible for the synthesis of an activated form of sulfate and are similar to the NodP and NodQ host specificity proteins of the bacterial symbiont *S. meliloti*.

In *S. meliloti*, two copies of the *nodPQ* operon are present. Both copies are involved in Nod factor sulfation but are not necessary for cysteine biosynthesis. Recently, in *S. meliloti* and in *Rhizobium tropici* CFN299, homologues of the *cysDN* (ATP sulfurylase) and *cysH* (APS reductase) genes were isolated, but no homologue of the *E. coli cysC* gene (APS kinase) could be identified (1, 23). Consequently, it was demonstrated that in *S. meliloti*, APS rather than PAPS is reduced for sulfite production during cysteine biosynthesis (1). Other members of the

* Corresponding author. Mailing address: Centre of Microbial and Plant Genetics, Kasteelpark Arenberg 20, B-3001 Heverlee, Belgium. Phone: 32 16 32 16 31. Fax: 32 16 32 19 63. E-mail: jozef.vanderleyden@agr.kuleuven.ac.be.

TABLE 1. Bacterial strains and plasmids

Strain or plasmid	Relevant characteristics	Reference or source
Sinorhizobium sp. strains		
BR816	Broad-host-range *Sinorhizobium* strain isolated from *Leucaena leucocephala*	16
FAJ1600	*cysD* mutant of BR816; Tc^r	This study
FAJ1604	*nodP1 nodQ2 cysD* triple mutant of BR816; Km^r Sp^r Tc^r	This study
CFNE205	*nodP1* mutant of BR816; Km^r	25
CFNE206	*nodQ2* deletion mutant of BR816; Sp^r	25
CFNE207	*nodP1 nodP2* double mutant of BR816; Km^r Sp^r	25
CFNE208	*nodP1 nodQ2* double mutant of BR816; Km^r Sp^r	T. Laeremans, unpublished results
Plasmids		
pBRE4.8	pUC19 carrying the BR816 *cysDN* genes; Ap^r	This study
pJQ200uc1	*B. subtilis sacB*-containing suicide vector; Gm^r	39
pHP45Ω-Tc	Vector containing Tc^r cassette	38
pUC18/19	Cloning vector; Ap^r	33

Rhizobiaceae, differing in their ability to incorporate sulfate in either a Nod factor or lipopolysaccharide, also preferentially reduce APS instead of PAPS for cysteine biosynthesis. This implies that APS reduction is not necessarily correlated with the presence of PAPS-dependent sulfurylation reactions for symbiosis, which is the case when functional *nodPQ* genes are present (1). Recently, Kopriva et al. (20) have described a phylogenetic classification of APS and PAPS reductase amino acid sequences (both annotated as CysH) from different organisms. The resulting sequence-based prediction of the substrate specificities of these enzymes was confirmed by Williams et al. (58), using genetic complementation experiments.

Sinorhizobium sp. strain BR816 (formerly *Rhizobium* sp. strain BR816) synthesizes Nod factors that are fully sulfated at the reducing terminal residue (50), as is the case for the narrow-host-range *S. meliloti* (26). The sulfate decoration on the Nod factors secreted by *S. meliloti* is essential for nodulation of alfalfa (40). Except for *S. meliloti*, it is still unclear whether rhizobia producing sulfated Nod factors use only the *nodPQ*-dependent PAPS pool as a source of activated sulfate for Nod factor sulfation, the housekeeping PAPS pool, or both (25). Previously, Laeremans et al. (25) demonstrated that *Sinorhizobium* sp. strain BR816 possesses two *nodPQ* copies. Although both copies are functional, as demonstrated by genetic complementation of an *R. tropici nodP* mutant, the double mutants did not show any detectable changes in the amount of sulfated Nod factors produced by this strain (25). It was suggested that in *Sinorhizobium* sp. strain BR816, in contrast to *S. meliloti*, a housekeeping locus as a third PAPS-producing locus could be involved in the sulfation of the Nod factors.

We have isolated the *cysDN* homologues of *Sinorhizobium* sp. strain BR816 and studied the role of this third PAPS-producing locus in relation to Nod factor synthesis. In addition, we were interested to know how the various forms of activated sulfate may be partitioned into the pathways for amino acid biosynthesis and sulfation or methylation of Nod factors and other compounds important during symbiosis. Furthermore, based on the analysis of the phylogenetic relationship among rhizobial ATP sulfurylases, we speculate on the possible origin and functionality of genes for sulfate activation.

MATERIALS AND METHODS

Bacterial strains and growth conditions. The bacterial strains and plasmids used in this study are listed in Table 1. *E. coli* strains were maintained on Luria-Bertani agar at 37°C and grown in Luria-Bertani broth (32). Rhizobial strains were maintained on yeast extract-mannitol medium (55) or on tryptone-yeast medium with added $CaCl_2$ (3) at 30°C. Antibiotics were added to the medium as needed at the following concentrations (micrograms per milliliter): ampicillin, 100; spectinomycin, 50; kanamycin, 50; and nalidixic acid, 31. Tetracycline was added to a final concentration of 1 μg/ml (for *Sinorhizobium* sp. strain BR816) or 10 μg/ml (for *E. coli*). Triparental conjugations and site-directed mutagenesis were done as previously described (31).

Nucleic acid manipulations and analysis. Isolation and cloning of plasmid DNA was performed as described previously (2, 42). Total genomic DNA of *Sinorhizobium* sp. strain BR816 was isolated by using a genomic DNA isolation kit (Gentra Systems) according to the manufacturer's instructions. DNA fragments were recovered from agarose gels by using the Nucleotrap kit (Macherey-Nagel). Southern blotting and hybridizations were carried out as previously described (25). Sequencing of DNA fragments cloned in the pUC18-pUC19 vectors was performed on an automated ALF sequencer with fluorescein-labeled universal and synthetic oligonucleotide primers (Amersham Pharmacia Biotech, Uppsala, Sweden). Database searches for similarity were performed with the BLAST software (National Center for Biotechnology Information, National Institutes of Health).

PCR was performed with *Taq* DNA polymerase (Boehringer, Mannheim, Germany) according to the manufacturer's protocol. For sequencing, the high-fidelity Platinum *Pfx* DNA polymerase (GIBCO-BRL, Life Technologies) was used according to the manufacturer's protocol.

To construct a genomic minilibrary, total genomic DNA from *Sinorhizobium* sp. strain BR816 was digested with *Eco*RI. DNA fragments ranging between 4 and 6 kb were recovered and ligated into the pUC19 cloning vector. Eight hundred Ap^r white colonies were picked up. Plasmid DNA was purified from 15 pools consisting of approximately 50 colonies, and efficient insertion of fragments of the desired size was confirmed. A 450-bp PCR fragment containing an internal part of *cysD* was used as a probe to screen the library.

Phylogenetic analysis of CysD homologues. The amino acid sequences of 19 CysD-like proteins, truncated to the same size as the shortest sequence (position 3 to 299 from the *S. meliloti* NodP1 sequence [gi14523565]) were aligned by using the ClustalW program (http://searchlauncher.bcm.tmc.edu/multi-align/multi-align.html). The construction of neighbor-joining trees (41) and bootstrap analysis of 1,000 resamples were performed by using the Treecon for Windows (1.3b) software package (53). In estimating evolutionary distances between amino acid sequences, we used the Poisson correction. Insertions and deletions were not taken into account. For constructing trees by the parsimony method, the PROTPARS program in the PHYLIP package was used (10). Again, bootstrap analysis of 1,000 resamples was performed.

Growth tests. Growth tests of *Sinorhizobium* sp. strain BR816 in sulfate-free liquid medium were carried out in acid minimal salts (AMS) medium (36) containing 1 mM $CaCl_2$ with sulfate salts replaced by equimolar amounts of alternative salts ($MgCl_2$, $ZnCl_2$, $MnCl_2$, and $CuCl_2$). Ammonium chloride (10

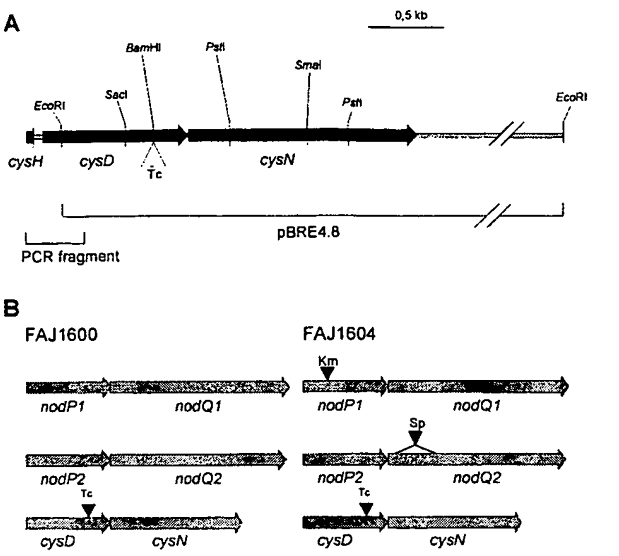


FIG. 1. (A) Physical and genetic maps of the BR816 *cysDN* region. The triangle indicates the position of the inserted Tc^r cassette in the mutants FAJ1600 and FAJ1604. (B) Schematic view of the constructed *Sinorhizobium* sp. strain BR816 mutants with mutations in the *nodPQ* genes and *cysDN* genes (see Table 1). Triangles indicate inserted antibiotic resistance cassettes.

mM) and mannitol (10 mM) were used as nitrogen and carbon sources, respectively. Sulfur compounds (sodium sulfate, sodium sulfite, L-cysteine, and L-methionine) were filter sterilized and added to the autoclaved medium at a concentration of 25 μM. When appropriate, cell cultures were induced with 500 nM apigenin. Cells of the strains tested were grown overnight in tryptone-yeast medium, washed twice in sulfate-free AMS medium, brought to an optical density of 0.4 (measured at 600 nm with a Perkin-Elmer lambda 2 spectrometer), and diluted 6,000-fold in sulfate-free AMS medium with the appropriate concentrations of filter-sterilized antibiotics, apigenin, and sulfur compounds. Bacteria were grown in microtiter plates (final volume, 300 μl) over a 4-day period, and cell growth was monitored automatically by measuring the optical density at 600 nm in BioscreenC (Labsystems) every 30 min. For each time point, the average optical density was calculated from five independent measurements.

Insertion mutagenesis. A *Sinorhizobium* sp. strain BR816 *cysD* single mutant and *nodP1 nodQ2 cysD* triple mutant were constructed as follows. To obtain the *cysD* single mutant, the 1.6-kb *Sma*I fragment of pBRE4.8 was ligated into the *Sma*I site of pJQ200uc1. This vector allows positive selection of double homologous recombinants on sucrose (10%)-containing medium due to the presence of the *Bacillus subtilis sacB* gene. The resulting plasmid was digested with *Bam*HI and then blunt-end ligated to the *Smu*I fragment containing the Ω-Tc^r cassette from pHP45Ω-Tc. This plasmid was conjugated to *Sinorhizobium* sp. strain BR816. Correct insertion of the Tc^r interposon was verified by Southern hybridization with the *cysD* gene and the Tc^r cassette as probes. In this way, the same construct was introduced in CFNE205 (*nodP1*), CFNE206 (*nodQ2*), CFNE207 (*nodP1 nodP2*), and CFNE208 (*nodP1 nodQ2*) (Table 1; Fig. 1). A *cysD* single mutant (FAJ1600) and a *nodP1 nodQ2 cysD* triple mutant (FAJ1604) were obtained and retained for further analysis.

Radioactive labeling of Nod metabolites and thin-layer chromatography (TLC) analysis. Nod factors were labeled by using the isotopes [^{14}C]acetate and [^{35}S]sulfate according to a slightly modified version of the protocol of Mergaert et al. (30), as previously described (25). For this experiment, Nod factors were purified from cells grown in sulfate-free AMS minimal medium supplemented with L-cysteine, as described for the growth tests.

Plant nodulation assay. Seeds of *Phaseolus vulgaris* cv. BAT477 were surface sterilized and germinated as described previously (56). Bean seedlings were planted in 250-ml flasks containing a nitrogen-free Snoeck medium agar slant (C. Snoeck, J. Vanderleyden, and E. Schrevens, submitted for publication) with KH_2PO_4 (7.49 mM), K_2SO_4 (0.43 mM), $CaCl_2$ (2.65 mM), $MgCl_2$ (1.75 mM), $MgSO_4$ (1.2 μM), FeNaEDTA (50.8 μM), $MnSO_4$ (35.2 μM), $CuSO_4$ (0.5 μM), $ZnSO_4$ (1.5 μM), H_3BO_3 (25 μM), and $(NH_4)_6Mo_7O_{24}$ (0.07 μM), with sulfate as the sole sulfur source unless otherwise stated. The seedlings were inoculated with approximately 10^6 bacteria per plant, from a diluted overnight culture that was washed twice with sulfate-free AMS medium. The plants were maintained in a growth chamber at 26°C (day) and 22°C (night) with a 12-h photoperiod. Plants were harvested after 3 weeks. Uninoculated control plants did not show any nodules or nodule-like structures. Ten plants per strain were tested in each experiment. Nitrogenase activity was determined by measuring the acetylene reduction activity of nodulated roots in closed vessels with a Hewlett-Packard 5890A gas chromatograph equipped with a PLOT fused silica column, with propane as an internal standard.

Data analysis. In all experiments, a randomized block design was used with 10 replicate blocks. Nodule number, nodule dry weight, and acetylene reduction activity were analyzed with the means and general linear model procedure (SAS Institute, Cary, N.C.). Comparison among the mean values obtained for each strain was made by Tukey's multiple-range test with a 95% confidence limit.

Nucleotide sequence accession number. Nucleotide sequence data were deposited in the GenBank database under accession number AJ505754.

RESULTS

Cloning and sequencing of a third PAPS-producing locus in *Sinorhizobium* sp. strain BR816. Previous work provided evidence for the presence of a putative third PAPS-producing locus in *Sinorhizobium* sp. strain BR816 on an approximately 4.8-kb *Eco*RI genomic DNA fragment (25). In order to clone this third copy of sulfate activation genes, a genomic minilibrary was constructed (see Materials and Methods), and a single positive clone, pBRE4.8, was obtained. Since the inserted genomic DNA region corresponding to the *cysD* gene was incomplete, the missing part of *cysD* was obtained by PCR with primers that were designed based on existing knowledge of the genomic organizations and DNA sequences of sulfate assimilation genes in other *Rhizobium* spp. (1, 23).

A physical map of the 4.8-kb *Eco*RI fragment and the upstream 442-bp PCR fragment was established (Fig. 1A), and the nucleotide sequence was determined. Similarity with an ATP sulfurylase encoded by the *cysD* and *cysN* genes of *S. meliloti*, *R. tropici* CFN299, and *E. coli* was found. Partial sequence similarity upstream of the *cysD* gene revealed the presence of a *cysH* homologue, encoding an APS or PAPS reductase, whereas no *cysC* homologue was found in the sequenced fragment. The same organization is found in *S. meliloti* and *R. tropici* (1, 23). It is likely that all three open reading frames are in a single operon, since no promoter consensus sequences or transcription termination signals were found in the intergenic *cysH-cysD* sequence of BR816. A similar situation was observed in *S. meliloti*, where two transcriptional start sites were identified, both upstream of the *cysH* homologue (1). In contrast, in *E. coli*, *cysH* does not form an operon with *cysDNC* (21). The *nodP* and *nodQ* homologues have a lower percent G+C content than the *cysD* and *cysN* homologues (data not shown), as observed for the *S. meliloti* genome (13).

The *Sinorhizobium* sp. strain BR816 *cysD* and *cysN* genes encode proteins of 317 and 498 amino acids, respectively. Strong conservation of amino acid residues was found with the respective CysD and CysN proteins of *S. meliloti* (96 and 91% identity, respectively), *R. tropici* (89 and 82% identity), and *E. coli* (68 and 52% identity). CysN contains the characteristic GTP-binding motif (GxxxxGK, DxxG, and NKxD) (7) and also an ITI motif, which is conserved among elongation factors (19). In comparison to the NodQ peptides, the deduced amino acid sequence of *cysN* lacks the carboxy-terminal part that corresponds to *E. coli* CysC. Therefore, no ATP-binding or PAPS-binding motifs were found. Similar observations were

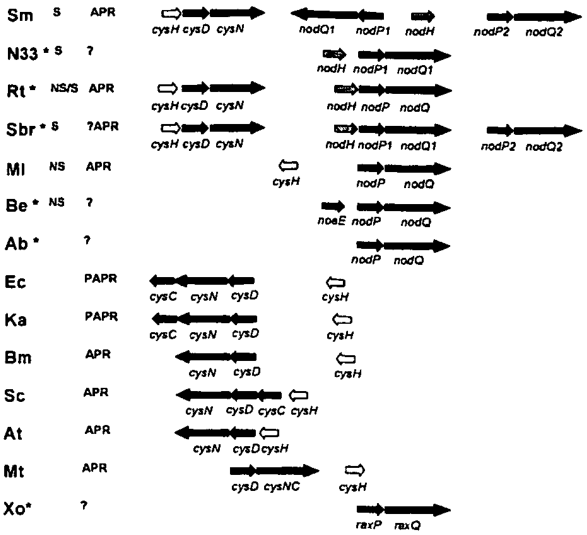


FIG. 2. Schematic representation of sulfate assimilation loci of selected strains for construction of a phylogenetic tree (Fig. 3). Abbreviations: Sm, *S. meliloti* (NodP1, gi14523565; NodP2, gi15140612; CysD, gi5911360); Sbr, *Sinorhizobium* sp. strain BR816 (NodP1, gi2148989; NodP2, gi27125923; CysD, gi24528409); Rt, *R. tropici* CFN299 (NodP, gi1280528; CysD, gi7387610); Bm, *Brucella melitensis* (CysD, gi17988038); N33, *Mesorhizobium* sp. strain N33 (NodP, gi1531624); Ml, *Mesorhizobium loti* (NodP, gi13476292); Be, *Bradyrhizobium elkanii* (NodP, gi14209498); Ab, *Azospirillum brasilense* (NodP, gi142424); Ec, *E. coli* (CysD, gi12517206); Ka, *Klebsiella aerogenes* (CysD, gi11992146); Xo, *X. oryzae* pv. oryzae (NodP, gi21105248); Mt, *Mycobacterium tuberculosis* (CysD, gi15608425); Sc, *Streptomyces coelicolor* (CysD, gi21224427); At, *Agrobacterium tumefaciens* (CysD, gi15155798). S, fully sulfated Nod factors; S/NS, mixture of sulfated and nonsulfated Nod factors; NS, nonsulfated Nod factors; APR, APS-reducing activity; PAPR, PAPS-reducing activity; ?APR, putative APR-reducing activity; ?, APS or PAPS reductase activity unknown; *, genome sequence not (fully) determined. Similar open reading frames are shaded identically. Note that *nodP1* of *S. meliloti* is located on megaplasmid 1, *nodP2* is on megaplasmid 2, and *cysHDN* is chromosomally located. *nodP1* of *Sinorhizobium* sp. strain BR816 is located on a megaplasmid, *nodP2* is on the symbiotic plasmid, and *cysHDN* is chromosomally located.

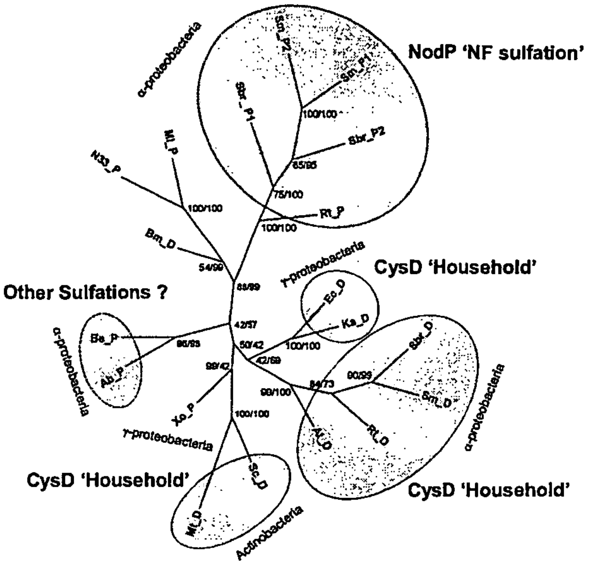


FIG. 3. Phylogenetic relationships among *cysD* gene products. The tree topology was inferred by using the protein parsimony method. Numbers represent the bootstrapping score (9) over 1,000 trials (parsimony/distance). The abbreviations of the species are as for Fig. 2.

made for *S. meliloti* and *R. tropici*. In summary, these data support the ATP sulfurylase activity of the putative proteins encoded by the isolated BR816 *cysDN* genes.

Phylogenetic analysis of CysD and CysN homologues. The BR816 CysD and CysN ATP sulfurylase subunits were compared through multiple-sequence alignment (ClustalW) with homologous ATP sulfurylases subunits retrieved from GenBank. The genomic organizations of the different sulfate assimilation loci of the strains selected for the phylogenetic analysis are schematically drawn in Fig. 2. Phylogenetic analysis of *cysD* and *nodP* gene products by the protein parsimony method resulted in a maximum-parsimony tree, as shown in Fig. 3. An identical tree topology could be inferred by using the neighbor-joining method (data not shown). Similar phylogenetic relationships could be deduced after construction of a phylogenetic dendrogram of CysN and NodQ protein sequences by using either the neighbor-joining method or protein parsimony analysis (data not shown).

It can be observed that the CysD and NodP ATP sulfurylase subunits of *Rhizobium* spp. producing sulfated Nod factors (Fig. 2), which have been shown to be involved in amino acid biosynthesis (1, 23) and Nod factor sulfation (24, 25, 47), respectively, cluster in two different groups (Fig. 3). The CysD protein of *Sinorhizobium* sp. strain BR816 clearly belongs to the protein cluster involved in biosynthesis of sulfur-containing amino acids, supporting its putative function.

Two other "household" clusters could be distinguished, i.e., the γ-*Proteobacteria* clade and the *Actinobacteria* clade. Interestingly, only one gene copy coding for a sulfate activation complex has been described, for *Mycobacterium tuberculosis* (*cysDNC*) (Fig. 2) (58). The sulfate assimilation pathway of *Mycobacterium tuberculosis* proceeds from sulfate through APS (catalyzed by CysDN), which is converted by APS reductase (CysH) in the first step toward cysteine and methionine. APS can also be converted to PAPS, through the action of the APS kinase CysC, and serves as a substrate for sulfotransferases that produce sulfolipids, which putatively function as virulence factors (58). Similarly, APS and PAPS pools are generated through the enzymatic activity of RaxP and RaxQ in *X. oryzae* pv. oryzae and are used for both cysteine synthesis and sulfation of avirulence effector molecules (48).

The CysD-homologous proteins of some members of the *Rhizobiaceae* (among which are *Mesorhizobium loti*, producing nonsulfated Nod factors [29, 34]; *Mesorhizobium* sp. strain N33, producing sulfated Nod factors [35]; and the pathogen *Brucella melitensis*) seem to belong to another cluster. However, these proteins are still more closely related to the NodP Nod factor sulfation cluster than to the CysD household cluster, as defined above. *Brucella melitensis* was previously shown to be genetically closely related to *Rhizobium* spp. (14). Intriguingly, the respective *Bradyrhizobium elkanii* and *Azospirillum brasilense* ATP sulfurylase subunits constitute a separate cluster (Fig. 3). The *nodPQ* genes of *B. elkanii* are situated within a gene cluster comprising genes for symbiotic functions (*fixGHIS* and *noeE*) as well as genes involved in rhizobitoxin biosynthesis (59). Since the *B. elkanii* Nod factors are not sulfated (4, 43), these genes do not function in Nod factor biosynthesis. The recently finished genome sequencing of the p90 plasmid of *A. brasilense* sheds new light on a possible function of its *nodPQ* copy, which is located within a region carrying genes involved in polysaccharide synthesis (E. Vanbleu and J. Vanderleyden, unpublished results). It was previously shown that *A. brasilense* does not synthesize Nod factors and that deletion of the *nodPQ* copy does not lead to auxotrophy (54). Therefore, it can be speculated that this cluster encompasses proteins belonging to a novel functionality group.

Growth characteristics of *Sinorhizobium* sp. strain BR816 *cysD* mutants under free-living conditions. To investigate the biochemical role of the isolated *cysDN* genes of BR816, the BR816 *cysD* gene was mutated (see Materials and Methods). First, the *cysD* mutants were tested for cysteine auxotrophy. In addition, we were interested to know whether a *cysD* mutation could be complemented by one or both *nodP* copies of *Sinorhizobium* sp. strain BR816. Growth of the wild type and various mutants with mutations in *nodPQ* and/or *cysDN* (FAJ1600, FAJ1604, CFNE205, CFNE206, CFNE207, and CFNE208) was examined in liquid sulfate-free AMS medium supplemented with various sulfur sources (see Materials and Methods). It could be demonstrated that the BR816 *nodPQ* single or double mutants (CFNE205, CFNE206, CFNE207, and CFNE208) exhibit growth patterns similar to that of the wild-type strain in minimal medium with sulfate as the sole sulfur source (data not shown). Therefore, it can be concluded that *nodPQ* mutants are not auxotrophs. Growth of the *cysD* mutant (FAJ1600) with sulfate as the sole sulfur source was clearly affected compared to that of the wild-type strain (Fig. 4A). FAJ1600 showed a prolonged lag phase, although its generation time in exponential growth phase did not markedly differ from that of the wild type. The *nodP1 nodQ2 cysD* triple mutant (FAJ1604) was completely impaired in growth (Fig. 4A). In the presence of sulfite, cysteine, or methionine, the growth of both mutants after 60 h was nearly restored to the wild-type level (Fig. 4B to D). This indicates that the *cysDN* genes are effectively involved in the biosynthesis of sulfur-containing amino acids, more specifically in the step of the sulfate assimilatory pathway just before the reduction of activated sulfate to sulfite. From this experiment we can conclude that knocking out the three sulfate activation systems (FAJ1604) in *Sinorhizobium* sp. strain BR816 leads to cysteine auxotrophy.

Interestingly, the growth characteristics of FAJ1600 showed a course similar to that of the wild type after a certain time interval. This demonstrates that the PAPS pool generated by the NodPQ sulfate activation complex is accessible for reduction by CysH and thus is available for the biosynthesis of sulfur-containing amino acids. The growth delay of FAJ1600 might indicate that CysH of *Sinorhizobium* sp. strain BR816 preferentially shows APS reductase activity rather than PAPS reductase activity toward the formation of sulfite. Moreover, the APS reductase activity of CysH has been recently confirmed in many rhizobial species (1, 20).

One should consider that (i) the growth curves of the wild-type and mutant strains were monitored under conditions in which no Nod factors are produced (no flavonoid induction) and (ii) *nodP2*, which is localized in the nodulation region on

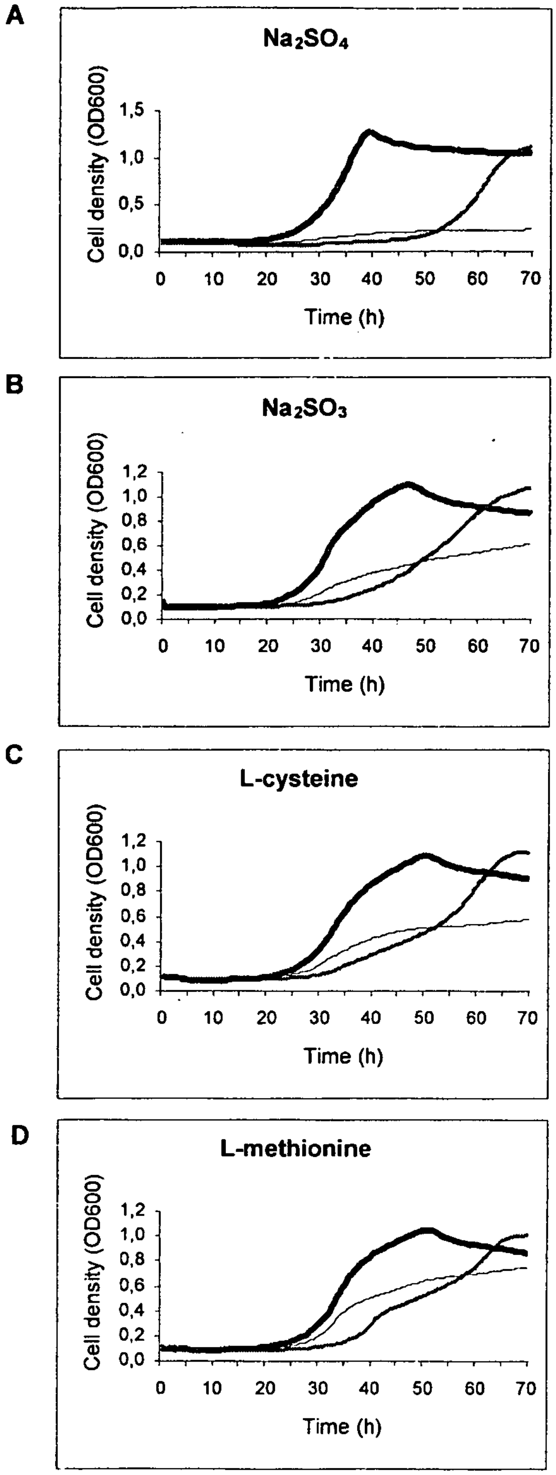


FIG. 4. Effects of various sulfur sources on cell growth of *Sinorhizobium* sp. strain BR816 wild-type and mutant strains determined by measuring optical density at 600 nm (OD600) in a BioscreenC instrument over a 4-day period. Thick black line, BR816; gray line, FAJ1600; thin gray line, FAJ1604). Cultures were grown at 30°C in sulfate-free AMS medium supplemented with sodium sulfate (A), sodium sulfite (B), L-cysteine (C), or L-methionine (D) at a concentration of 25 μM. Each experiment was conducted three times. Results from one experiment are shown.

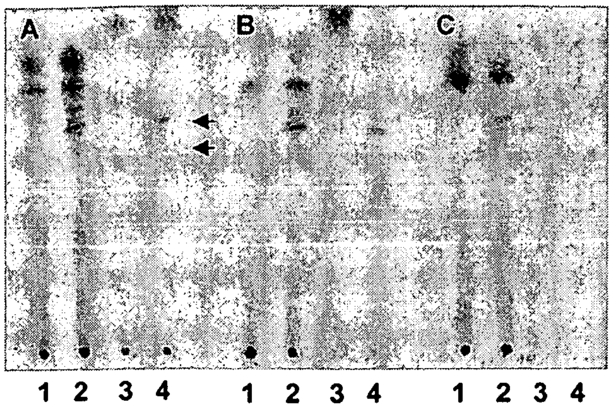


FIG. 5. Autoradiogram of a reverse-phase TLC profile of butanol extracts of radioactively labeled *Sinorhizobium* sp. strain BR816 (A), FAJ1600 (B), and FAJ1604 (C). Lanes 1 and 2, ^{14}C labeling; lanes 3 and 4, ^{35}S labeling. Lanes 1 and 3, noninduced; lanes 2 and 4, apigenin induced. Spots representing sulfated Nod factors are indicated with arrows.

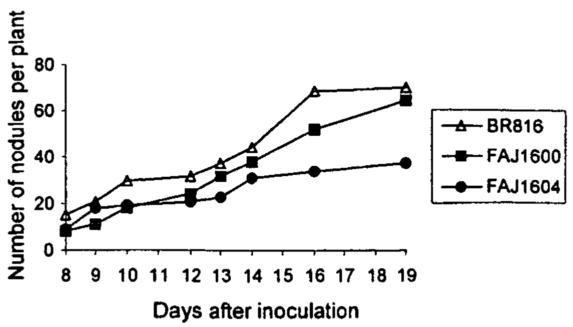


FIG. 6. Nodulation kinetics of *P. vulgaris* BAT477 inoculated with *Sinorhizobium* sp. strain BR816 wild-type and mutant strains. Two independent experiments were set up, and the results of one experiment are shown.

the symbiotic plasmid, probably is *nod* box dependent and thus not expressed (49). Therefore, to investigate whether the simultaneous production of sulfated Nod factors affects growth characteristics of the *cysDN* mutant strains, similar growth tests were performed in the presence of the *nod* gene inducer apigenin and with sulfate as the sole sulfur source. In this case, similar growth courses were obtained for FAJ1600 and FAJ1604 compared to the wild type (data not shown). This implies that at least the expressed *nodPQ* copy can complement and is sufficient for growth of FAJ1600 in minimal medium with sulfate as the sole sulfur source. The use of higher concentrations of inducer did not have a significant effect on the growth curves of the strains tested.

Nod factor sulfation pattern of *Sinorhizobium* sp. strain BR816 *cysD* mutants. Since the available *nodPQ* single and double mutants of *Sinorhizobium* sp. strain BR816 (CFNE205, CFNE206, CFNE207, and CFNE208 [Table 1]) were not auxotrophic and still produced sulfated Nod factors, Laeremans et al. (25) speculated that the housekeeping *cysDN*(*C*) genes can complement mutations in genes responsible for Nod factor sulfation. In order to determine to what level the Nod factors produced by the wild-type strain and the mutant strains FAJ1600 and FAJ1604 were still sulfated, apigenin-induced cell cultures, grown in liquid sulfate-free AMS medium supplemented with cysteine, were labeled with [^{14}C]acetate or [^{35}S]sulfate, and butanol extracts of the cell cultures were analyzed by reverse-phase TLC. Separation of the BR816 Nod factors revealed the presence of apigenin-induced spots on the chromatogram, corresponding to the Nod factors of BR816 (Fig. 5). The triple mutant FAJ1604 no longer produced sulfated Nod factors, which is in clear contrast with the sulfated Nod factor pattern of both the wild-type strain and FAJ1600 (Fig. 5). These results indicate that an activated sulfate source needed for Nod factor sulfation is no longer present. It can be concluded that the *cysDN* sulfate assimilation locus does provide active sulfate for NF sulfation.

Symbiotic phenotype of *cysD* mutants. The *Sinorhizobium* sp. strain BR816 *cysD* mutants were tested for their ability to nodulate common bean (*P. vulgaris* cv. BAT477) and to fix nitrogen. No significant differences in the kinetics of appearance of the first nodules were observed (Fig. 6). However, FAJ1600 (*cysD*) as well as FAJ1604 (*nodP1 nodQ2 cysD*) showed a decreased nodule number per plant over time, but only for FAJ1604 was this difference significant at the 95% level (Tukey's test). Morphologically, the nodules of both mutant strains were generally smaller with apparently less leghemoglobin present (as judged by the absence of pink color).

To study the nitrogen fixation capacity of the nodulated roots, the acetylene reduction activity was measured. The acetylene reduction activity of 21-day-old nodules induced by FAJ1600 or FAJ1604 was significantly lower than that for the wild-type strain ($P < 0.05$; Tukey's test) (data not shown). When methionine was added to the plant nutrient solution, the nitrogen fixation per plant was restored to wild-type levels. Interestingly, supplementation with methionine resulted in an overall higher nitrogen fixation capacity of *P. vulgaris* cv. BAT477 inoculated with *Sinorhizobium* sp. strain BR816 (data not shown).

DISCUSSION

In this study, a third APS-producing locus of the broad-host-range strain *Sinorhizobium* sp. strain BR816 was isolated. The nucleotide sequence of this region was determined, and based on homology searches, *cysD* and *cysN* were identified. Like in *S. meliloti*, no *cysC* homologue could be isolated downstream from *cysDN*. This is an indication that, like in other rhizobia, APS rather than PAPS is reduced to sulfite for cysteine biosynthesis (1). The highest similarity was found with the *cysDN* homologues in *S. meliloti*, supporting the close phylogenetic relationship between *S. meliloti* and *Sinorhizobium* sp. strain BR816 (15). Phylogenetic analysis revealed that CysD does not cluster with NodP1 and NodP2. The two BR816 NodP proteins are closely related and could have originated from a recent gene duplication, as was proposed for the NodP proteins of *S. meliloti* (13). Within the α-*Proteobacteria* clade, two clusters of proteins are clearly functionally distinguished and were designated NodP Nod factor sulfation and CysD household. It has been demonstrated that the *nodPQ* genes are also required for sulfation of *S. meliloti* lipopolysaccharide, proving a dual functionality of members of the NodP Nod factor sulfation cluster

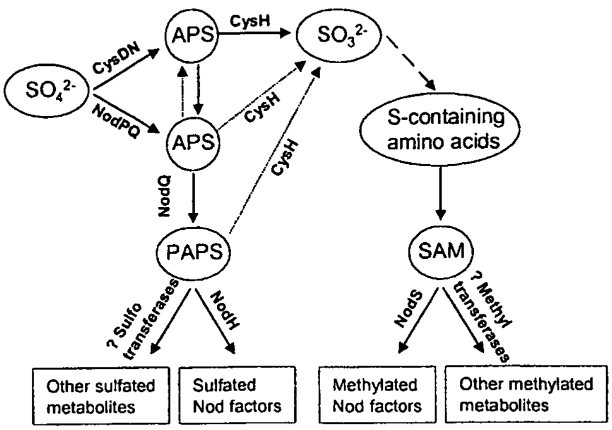


FIG. 7. Schematic representation of the distribution of APS and PAPS for sulfation and methylation processes in *Sinorhizobium* sp. strain BR816. Dotted arrows indicate possible but less favorable enzyme activity.

(5, 17). A potential new NodP-like protein cluster is proposed, comprising proteins involved in sulfate activation for sulfation of compounds that are yet unknown but which could be important during symbiosis. Other closely related CysD and NodP homologous do not fit into a specific functionality group, since these proteins are involved either in sulfation of amino acids (*M. loti* and *B. melitensis*) or in sulfation of Nod factors (*Mesorhizobium* sp. strain N33). It should be noted that within the γ-*Proteobacteria* clade and the *Actinobacteria* clade, only one copy of genes encoding sulfate activating enzymes is present, which seems to be involved in biosynthesis of sulfur-containing amino acids as well as sulfation of other macromolecules.

We examined the effect of a *cysD* mutation under free-living conditions in a wild-type chromosomal background and in a *nodP1 nodQ2* double mutant background. The levels of Nod factor sulfation (Fig. 5) as well as the growth characteristics (Fig. 4) of the different mutants were determined. In this study, we could demonstrate that the household CysDN sulfate activation locus of BR816 can additionally ensure Nod factor sulfation, whereas the symbiotic (P)APS pool, generated by the *nodPQ* sulfate activation complexes, can be engaged for sulfation of amino acids. Figure 7 shows a model of how the various forms of activated sulfate in *Sinorhizobium* sp. strain BR816 may be partitioned into the pathways for amino acid biosynthesis and sulfation of Nod factors and other compounds that might be important during symbiosis. The *cysDN*-dependent APS pool supplies activated sulfate that is subsequently reduced to form sulfite by the CysH APS reductase. Sulfite is further reduced to sulfide, which is then incorporated into the cysteine and methionine biosynthesis pathway. Our data suggest that the symbiotic APS and/or PAPS pool, created by the *nodPQ*-dependent sulfate activation step, can also be used by CysH (in a less efficient manner) for the biosynthesis of sulfur-containing amino acids, when needed. Moreover, both household and symbiotic APS pools can be mutually exchanged. In *S. meliloti*, the *nodPQ*- and *cysDN*-encoded sulfate activation systems cannot substitute for each other (46, 47).

Why would *Sinorhizobium* sp. strain BR816 possess three functional sulfate activation systems for Nod factor sulfation? Besides the use of activated sulfate for the biosynthesis of sulfur-containing amino acids and sulfation of Nod factors, (P)APS is needed for Nod factor methylation (37). Introduction of the *S. meliloti nodPQ* genes into *R. tropici* resulted in a decreased rate of *R. tropici* Nod factor methylation, while all *R. tropici* Nod factor backbones were sulfated. Waelkens et al. (57) showed that methylation of Nod factors is required for nodulation of bean. In *Sinorhizobium* sp. strain BR816, the three operational sulfate-activating systems could play an important role in maintaining substitutions of bacterial determinants for symbiosis.

An *R. tropici nodPQ* mutant (producing drastically reduced amounts of sulfated Nod factors) and an *R. tropici nodH* mutant (producing nonsulfated Nod factors) still activate the signaling cascade for emergence of effective nodules on *P. vulgaris* roots (12, 24). For bean plants, the sulfate moiety of the Nod factor was shown to be involved in the efficiency of nodule formation but appears not to be essential (11, 22). The effects of the *cysD* mutant FAJ1600 and the *nodP1 nodQ2 cysD* triple mutant FAJ1604 on bean symbiosis were seen mainly in the reduction of nodule number per plant. Since under free-living conditions, a *cysDN*-dependent biosynthesis of sulfur-containing amino acids is essential to allow optimal growth of *Sinorhizobium* sp. strain BR816 with sulfate as the sole sulfur source, bacterial growth defects are likely the main reason for the decreased nitrogen fixation of bean plants inoculated with the mutants FAJ1600 and FAJ1604. These defects can be restored by the addition of methionine to the plant nutrient solution. We propose that at the early stages of the nodulation, the plant root exudates of the germinated seedlings provide enough sources of organic sulfur to allow bacterial growth. However, a shortage of an organic sulfur source like methionine impairs bacterial growth inside the plant. Inoculation experiments with a *Rhizobium etli metZ* (*O*-succinylhomoserine sulfhydrylase for methionine biosynthesis) (51) mutant on bean plants resulted in the formation of ineffective (Nod$^+$ Fix$^-$) nodules, which suggested that root cells do not supply the inoculant bacteria with enough methionine. The fact that supplemented methionine resulted in an overall higher nitrogen fixation capacity of *P. vulgaris* BAT477 inoculated with BR816 strains supports this hypothesis. In contrast to our observations, an *R. etli cysG* (siroheme synthetase for cysteine biosynthesis) mutant, which is able to induce the formation of effective nodules (Nod$^+$ Fix$^+$) on the roots of common bean, seems to dispose of an organic sulfur source like cysteine or glutathione to allow growth inside the plant (52).

How can the strictly separated symbiotic and endogenous (P)APS pools in *S. meliloti* versus the complementary (P)APS pools in *Sinorhizobium* sp. strain BR816 be explained? Presumably, the *nodPQ* genes arose in ancestral rhizobial strains through duplications of the endogenous *cysDNC* genes. Later, these *nodPQ* genes evolved toward more specialized symbiotic genes, whereas the endogenous *cysC* gene, encoding the APS kinase, was apparently lost during evolution. At this stage, complementation between both PAPS pools was still possible (the case of *Sinorhizobium* sp. strain BR816). Then, the genetic separation of the two sulfate-activating systems could have further evolved into two more efficient and energy-saving separate enzymatic multienzyme complexes (the case of *S. meliloti*).

ACKNOWLEDGMENTS

C.S. is the recipient of a postdoctoral fellowship from the Research Council of the Katholieke Universiteit Leuven.

We thank A. Parret and B. Dombrecht for their help with the construction of the phylogenetic tree.

REFERENCES

1. **Abola, A. P., M. Willits, R. Wang, and S. Long.** 1999. Reduction of adenosine-5′-phosphosulfate instead of 3′-phosphoadenosine-5′phosphosulfate in cysteine biosynthesis by *Rhizobium meliloti* and other members of the family *Rhizobiaceae*. J. Bacteriol. **181:**5280–5287.
2. **Ausubel, F. M., R. Brent, R. E. Kingston, D. D. Moore, J. G. Seidman, J. A. Smith, and K. Struhl.** 1987. Current protocols in molecular biology. John Wiley and Sons, New York, N.Y.
3. **Beringer, J. E.** 1974. R-factor transfer in Rhizobium leguminosarum. J. Gen. Microbiol. **120:**421–429.
4. **Carlson, R. W., J. Sanjuan, U. Ramadas Bhat, J. Glushka, H. P. Spaink, A. H. Wijfjes, A. A. N. van Brussel, T. J. W. Stokkermans, N. Kent Peters, and G. Stacey.** 1993. The structures and biological activities of the lipo-chitooligosaccharide nodulation signals produced by type I and II strains of *Bradyrhizobium japonicum*. J. Biol. Chem. **268:**18372–18381.
5. **Cerdergren, R., J. Lee, K. Ross, and R. Hollingsworth.** 1995. Common links in the structure and cellular localization of *Rhizobium* chitolipooligosaccharides and general *Rhizobium* membrane phospholipid and glycolipid components. Biochemistry **34:**4467–4477.
6. **Cloutier, J., S. Laberge, Y. Castonguay, and H. Antoun.** 1996. Characterization and mutational analysis of *nodHPQ* genes of *Rhizobium* sp. N33. Mol. Plant-Microbe Interact. **9:**720–728.
7. **Dever, T. E., M. J. Glynias, and W. C. Merrick.** 1987. GTP-binding domain: three consensus sequence elements with distinct spacing. Proc. Natl. Acad. Sci. USA **84:**1814–1818.
8. **Ehrhardt, D. W., E. M. Atkinson, K. F. Faull, D. I. Freedberg, D. P. Sutherlin, R. Armstrong, and S. R. Long.** 1995. In vitro sulfotransferase activity of NodH, a nodulation protein of *Rhizobium meliloti* required for host-specific nodulation. J. Bacteriol. **177:**6237–6245.
9. **Felsenstein, J.** 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39:**783–791.
10. **Felsenstein, J.** 1993. PHYLIP (Phylogeny Inference Package), version 3.5c. Department of Genetics, University of Washington, Seattle.
11. **Fernández-López, M., W. D'Haeze, M. Van Montagu, and M. Holsters.** 1998. Changes in glycosylation pattern at the reducing end of azorhizobial Nod factors affect nodulation efficiency. FEMS Microbiol. Lett. **158:**237–242.
12. **Folch-Mallol, J. L., S. Marroqui, C. Sousa, H. Manyani, I. M. Lopez Lara, K. M. G. M. vander Drift, J. Haverkamp, C. Quinto, A. GilSerrano, J. Thomas-Oates, H. P. Spaink, and M. Megias.** 1996. Characterization of *Rhizobium tropici* CIAT899 nodulation factors: the role of *nodH* and *nodPQ* genes in their sulfation. Mol. Plant-Microbe Interact. **9:**151–163.
13. **Galibert, F., T. M. Finan, S. R. Long, A. Puhler, P. Abola, F. Ampe, F. Barloy-Hubler, M. J. Barnett, A. Becker, P. Boistard, G. Bothe, M. Boutry, L. Bowser, J. Buhrmester, E. Cadieu, D. Capela, P. Chain, A. Cowie, R. W. Davis, S. Dreano, N. A. Federspiel, R. F. Fisher, S. Gloux, T. Godrie, A. Goffeau, B. Golding, J. Gouzy, M. Gurjal, I. Hernandez-Lucas, A. Hong, L. Huizar, R. W. Hyman, T. Jones, D. Kahn, M. L. Kahn, S. Kalman, D. H. Keating, E. Kiss, C. Komp, V. Lelaure, D. Masuy, C. Palm, M. C. Peck, T. M. Pohl, D. Portetelle, B. Purnelle, U. Ramsperger, R. Surzycki, P. Thebault, M. Vandenbol, F. J. Vorholter, S. Weidner, D. H. Wells, K. Wong, K. C. Yeh, and J. Batut.** 2001. The composite genome of the legume symbiont *Sinorhizobium meliloti*. Science **293:**668–672.
14. **Gándara, B., A. López Merino, M. A. Rogel, and E. Martínez-Romero.** 2001. Limited genetic diversity of *Brucella* spp. J. Clin. Microbiol. **39:**235–240.
15. **Hernández-Lucas, I., L. Segovia, E. Martínez-Romero, and S. G. Pueppke.** 1995. Phylogenetic relationships and host range of *Rhizobium* spp. that nodulate *Phaseolus vulgaris* L. Appl. Environ. Microbiol. **61:**2775–2779.
16. **Hungria, M., A. A. Franco, and J. L. Sprent.** 1993. New sources of high-temperature tolerant rhizobia for *Phaseolus vulgaris* L. Plant Soil **149:**103–109.
17. **Keating, D. H., M. G. Willits, and S. R. Long.** 2002. A *Sinorhizobium meliloti* lipopolysaccharide mutant altered in cell surface sulfation. J. Bacteriol. **184:** 6681–6689.
18. **Kertesz, M. A.** 2001. Bacterial transporters for sulfate and organosulfur compounds. Res. Microbiol. **152:**279–290.
19. **Kohno, K., T. Uchida, H. Ohkubo, S. Nakanishi, T. Nakanishi, T. Fukui, E. Ohtsuka, M. Ikehara, and Y. Okada.** 1986. Amino acid sequence of mammalian elongation factor 2 deduced from the cDNA sequence: homology with GTP-binding proteins. Proc. Natl. Acad. Sci. USA **83:**4978–4982.
20. **Kopriva, S., T. Büchcr, G. Fritz, M. Suter, R. Benda, V. Schünemann, A. Koprivova, P. Schürmann, A. X. Trautwein, P. M. H. Kroneck, and C. Brunold.** 2002. The presence of an iron-sulfur cluster in adenosine 5′-phosphosulfate reductase separates organisms utilizing adenosine 5′-phosphosulfate and phosphoadenosine 5′-phosphosulfate for sulfate assimilation. J. Biol. Chem. **277:**21786–21791.
21. **Kredich, N. M.** 1996. Biosynthesis of cysteine, p. 514–527. *In* F. C. Neidhardt, R. Curtiss III, J. L. Ingraham, E. C. C. Lin, K. B. Low, B. Magasanik, W. S. Reznikoff, M. Riley, M. Schaechter, and H. E. Umbarger (ed.), *Escherichia coli* and *Salmonella*: cellular and molecular biology. American Society for Microbiology, Washington D.C.
22. **Laeremans, T., C. Snoeck, J. Mariën, C. Verreth, E. Martínez-Romero, J.-C. Promé, and J. Vanderleyden.** 1999. *Phaseolus vulgaris* (L.) recognizes *Azorhizobium caulinodans* Nod factors with a variety of chemical substituents. Mol. Plant-Microbe Interact. **12:**820–824.
23. **Laeremans, T., E. Martínez-Romero, and J. Vanderleyden.** 1998. Isolation and sequencing of a second *Rhizobium tropici* CNF299 genetic locus that contains genes homologous to amino acid sulphate activation genes. DNA Sequence **9:**65–70.
24. **Laeremans, T., I. Caluwaerts, C. Verreth, M. A. Rogel, J. Vanderleyden, and E. Martinez-Romero.** 1996. Isolation and characterization of the *Rhizobium tropici* Nod factor sulfation genes. Mol. Plant-Microbe Interact. **9:**492–500.
25. **Laeremans, T., N. Coolsaet, C. Verreth, C. Snoeck, N. Hellings, J. Vanderleyden, and E. Martínez-Romero.** 1997. Functional redundancy of genes for sulfate activation enzymes in *Rhizobium* sp. BR816. Microbiology **143:**3933–3942.
26. **Lerouge, P., P. Roche, C. Faucher, F. Maillet, G. Truchet, J.-C. Promé, and J. Dénarié.** 1990. Symbiotic host-specificity of *Rhizobium meliloti* is determined by a sulphated and acetylated glucosamine oligosaccharide signal. Nature **344:**781–784.
27. **Leyh, T. S., J. C. Taylor, and G. D. Markham.** 1988. The sulfate activation locus of *Escherichia coli* K12: cloning, genetic and enzymatic characterization. J. Biol. Chem. **263:**2409–2416.
28. **Leyh, T. S., T. F. Vogt, and Y. Suo.** 1992. The DNA sequence of the sulfate activation locus from *Escherichia coli* K-12. J. Biol. Chem. **267:**10405–10410.
29. **López-Lara, I. M., J. D. Van den Bergh, J. E. Thomas-Oates, J. Glushka, B. Lugtenberg, and H. P. Spaink.** 1995. Structural identification of the lipo-chitin oligosaccharide nodulation signals of *Rhizobium loti*. Mol. Microbiol. **15:**627–638.
30. **Mergaert, P., M. Van Montagu, J.-C. Promé, and M. Holsters.** 1993. Three unusual modifications, a *D*-arabinosyl, an *N*-methyl, and a carbamoyl group, are present on the Nod factors of *Azorhizobium caulinodans* strain ORS571. Proc. Natl. Acad. Sci. USA **90:**1551–1555.
31. **Michiels, J., M. Moris, B. Dombrecht, C. Verreth, and J. Vanderleyden.** 1998. Differential regulation of *Rhizobium etli rpoN2* gene expression during symbiosis and free-living growth. J. Bacteriol. **180:**3620–3628.
32. **Miller, J. H.** 1972. Experiments in molecular genetics. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
33. **Norrander, J., T. Kempe, and J. Messing.** 1983. Construction of improved M13 vectors using oligodeoxynucleotide-directed mutagenesis. Gene. **26:** 101–106.
34. **Olsthoorn, M. M. A., I. M. López-Lara, B. O. Petersen, K. Bock, J. Haverkamp, H. P. Spaink, and J. E. Thomas-Oates.** 1998. Novel branched Nod factor structure results from α-(1→3) fucosyl transferase activity: the major lipo-chitin oligosaccharides from *Mesorhizobium loti* strain NZP2213 bear an α-(1→3) fucosyl substituent on a non-terminal backbone residue. Biochemistry **37:**9024–9032.
35. **Poinsot, V., E. Belanger, S. Laberge, G. P. Yang, H. Antoun, J. Cloutier, M. Treilhou, J. Denarie, J.-C. Prome, and F. Debellé.** 2001. Unusual methyl-branched α,β-unsaturated acyl chain substitutions in the Nod factors of an arctic *Rhizobium*, *Mesorhizobium* sp. strain N33 (*Oxytropis arctobia)*. J. Bacteriol. **183:**3721–3728.
36. **Poole, P. S., N. A. Schofield, C. J. Reid, E. M. Drew, and D. L. Walshaw.** 1994. Identification of chromosomal genes located downstream of *dctD* that affect the requirement for calcium and the lipopolysaccharide layer of *Rhizobium leguminosarum*. Microbiology **140:**2797–2809.
37. **Poupot, R., E. Martinez-Romero, N. Gautier, and J.-C. Promé.** 1995. Wild-type *Rhizobium etli*, a bean symbiont, produces acetyl-fucosylated, N-methylated, and carbamoylated nodulation factors. J. Biol. Biochem. **270:**6050–6055.
38. **Prentki, P., and H. M. Kirsch.** 1984. *In vitro* insertional mutagenesis with a selectable DNA fragment. Gene **29:**303–313.
39. **Quandt, J., and M. F. Hynes.** 1993. Versatile suicide vectors which allow direct selection for gene replacement in Gram-negative bacteria. Gene **127:** 15–21.
40. **Roche, P., F. Debellé, F. Maillet, P. Lerouge, C. Faucher, G. Truchet, J. Dénarié, and J.-C. Promé.** 1991. Molecular basis of symbiotic host specificity in *Rhizobium meliloti*: *nodH* and *nodPQ* genes encode the sulfation of lipo-oligosaccharide signals. Cell **67:**1131–1143.
41. **Saitou, N., and M. Nei.** 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:**406–425.
42. **Sambrook, J., E. F. Fritsch, and T. Maniatis.** 1989. Molecular cloning: a

laboratory manual, 2nd ed. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.

43. **Sanjuan, J., R. W. Carlson, H. P. Spaink, U. R. Bhat, W. M. Barbour, J. Glushka, and G. Stacey.** 1992. A 2-O-methylfucose moiety is present in the lipo-oligosaccharide nodulation signal of *Bradyrhizobium japonicum*. Proc. Natl. Acad. Sci. USA **89:**8789–8793.
44. **Schultze, M., C. Staehelin, H. Röhrig, M. John, J. Schmidt, and A. Kondorosi.** 1995. *In vitro* sulfotransferase activity of *Rhizobium meliloti* NodH protein: lipochito-oligosaccharide nodulation signals are sulfated after synthesis of the core structure. Proc. Natl. Acad. Sci. USA **92:**2706–2709.
45. **Schwedock, J., and S. Long.** 1990. ATP sulfurylase activity of the *nodP* and *nodQ* gene products of *Rhizobium meliloti*. Nature **348:**644–647.
46. **Schwedock, J., and S. Long.** 1992. *Rhizobium meliloti* genes involved in sulfate activation: two copies of *nodPQ* and a new locus, *saa*. Genetics **132:**899–909.
47. **Schwedock, J., and S. R. Long.** 1994. *Rhizobium meliloti* NodP and NodQ form a multifunctional sulfate-activating complex requiring GTP for activity. J. Bacteriol. **176:**7055–7064.
48. **Shen, Y., P. Sharma, F. G da Silva, and P. Ronald.** 2002. The *Xanthomonas oryzae* pv. *oryzae raxP* and *raxQ* genes encode an ATP sulphurylase and adenosine-5′-phosphosulphate kinase that are required for AvrXa21 avirulence activity. Mol. Microbiol. **44:**37–48.
49. **Snoeck, C.** 2001. Host specificity determinants of *Sinorhizobium* sp. BR816 for early signaling in symbiotic interactions. Ph.D. thesis. Katholieke Universiteit Leuven, Leuven, Belgium.
50. **Snoeck, C., E. Luyten, V. Poinsot, A. Savagnac, J. Vanderleyden, and J.-C. Promé.** 2001. *Rhizobium* sp. BR816 produces a complex mixture of known and novel lipo-chitooligosaccharide molecules. Mol. Plant-Microbe Interact. **14:**678–684.
51. **Taté, R., A. Riccio, E. Caputo, M. Iaccarino, and E. J. Patriarca.** 1999. The *Rhizobium etli metZ* gene is essential for methionine biosynthesis and nodulation of *Phaseolus vulgaris*. Mol. Plant-Microbe Interact. **12:**24–34.
52. **Taté, R., A. Riccio, M. Iaccarino, and E. J. Patriarca.** 1997. A *cysG* mutant strain of *Rhizobium etli* pleiotropically defective in sulfate and nitrate assimilation. J. Bacteriol. **179:**7343–7350.
53. **Van de Peer, Y., and R. De Wachter.** 1994. TREECON for Windows: a software package for the construction and drawing of evolutionary trees for the Microsoft Windows environment. Comput. Appl. Biosci. **10:**569–570.
54. **Vielle, C., and C. Elmerich.** 1990. Characterization of two *Azospirillum brasilense* Sp7 plasmid genes homologous to *Rhizobium meliloti nodPQ*. Mol. Plant-Microbe Interact. **3:**389–400.
55. **Vincent, J. M.** 1970. A manual for the practical study of the root-nodule bacteria. Blackwell Scientific Publications, Oxford, United Kingdom.
56. **Vlassak, K. M., E. Luyten, C. Verreth, P. van Rhijn, T. Bisseling, and J. Vanderleyden.** 1998. The *Rhizobium* sp. BR816 *nodO* gene can function as a determinant for nodulation of *Leucaena leucocephala*, *Phaseolus vulgaris* and *Trifolium repens* by a diversity of *Rhizobium* spp. Mol. Plant-Microbe Interact. **5:**383–392.
57. **Waelkens, F., T. Voets, K. Vlassak, J. Vanderleyden, and P. van Rhijn.** 1995. The *nodS* gene of *Rhizobium tropici* strain CIAT899 is necessary for nodulation on *Phaseolus vulgaris* and on *Leucaena leucocephala*. Mol. Plant-Microbe Interact. **8:**147–154.
58. **Williams, S. J., R. H. Senaratne, J. D. Mougous, L. W. Riley, and C. R. Bertozzi.** 2002. 5′-Adenosinephosphosulfate lies at a metabolic branchpoint in mycobacteria. J. Biol. Chem. **277:**32606–32615.
59. **Yasuta, T., S. Okazaki, H. Mitsui, K. Yuhashi, H. Ezura, and K. Minamisawa.** 2001. DNA sequence and mutational analysis of rhizobitoxine biosynthesis genes in *Bradyrhizobium elkanii*. Appl. Environ. Microbiol. **67:**4999–5009.

APPENDIX L

ExPASy Home page Site Map Search ExPASy Contact us PROSITE

Search PROSITE for Go Clear

NiceSite View of: PS00629

General information about the entry	
Entry name	**IMP_1**
Accession number	**PS00629**
Entry type	PATTERN
Date	JUN-1992 (CREATED); APR-2006 (DATA UPDATE); JUL-2006 (INFO UP
PROSITE documentation	PDOC00547
Name and characterization of the entry	
Description	Inositol monophosphatase family signature 1.
Pattern	[FWV]-x(0,1)-[LIVM]-D-P-[LIVM]-D-[SG]-[ST]-x(2)-[FYA]-x(0,1)-[HKRNSTY

Numerical results

- UniProtKB/Swiss-Prot release number: **50.4**, total number of sequence entries in that r **230133**.
- Total number of hits in UniProtKB/Swiss-Prot: **75 hits in 75 different sequences**
- Number of hits on proteins that are known to belong to the set under consideration: **74 different sequences**
- Number of hits on proteins that could potentially belong to the set under consideration: **different sequences**
- Number of false hits (on unrelated proteins): **0 hits in 0 different sequences**
- Number of known missed hits: **5**
- Number of partial sequences which belong to the set under consideration, but which ar pattern or profile because they are partial (fragment) sequences: **1**
- Precision (true hits / (true hits + false positives)): **100.00 %**
- Recall (true hits / (true hits + false negatives)): **93.67 %**

Comments

- Taxonomic range: **Archaebacteria, Eukaryotes, Prokaryotes (Bacteria)**
- Maximum known number of repetitions of the pattern in a single protein: **1**
- `Interesting' site in the pattern: **4,metal(?)**
- `Interesting' site in the pattern: **9,metal(?)**
- VERSION: **2**

Cross-references

True positive hits:

```
BPNT1_HUMAN (O95861), BPNT1_MOUSE (Q9Z0S1), BPNT1_RAT
CYSQ_ACTAC  (P70714), CYSQ_BUCAI  (P57624), CYSQ_ECO57
CYSQ_ECOL6  (Q8FAG5), CYSQ_ECOLI  (P22255), CYSQ_HAEIN
CYSQ_MYCBO  (P65164), CYSQ_MYCLE  (P46726), CYSQ_MYCTU
```

UniProtKB/Swiss-Prot

```
CYSQ_SALTI   (Q8Z153), CYSQ_SALTY   (P26264), CYSQ_SHIFL
DPNP1_ARATH  (Q42546), DPNP2_ARATH  (O49623), DPNP4_ARATH
DPNPH_ARATH  (Q38945), DPNPM_ARATH  (Q9M0Y6), DPNP_ORYSA
DPNP_SCHPO   (O94505), HAL21_CANAL  (P46594), HAL22_CANAL
HAL2_YEAST   (P32179), IMP1_LYCES   (P54926), IMP2_LYCES
IMP3_LYCES   (P54928), IMPA1_BOVIN  (P20456), IMPA1_CAEEL
IMPA1_HUMAN  (P29218), IMPA1_MOUSE  (O55023), IMPA1_PIG
IMPA1_PONPY  (Q5R4X0), IMPA1_RAT    (P97697), IMPA1_XENLA
IMPA2_HUMAN  (O14732), IMPA2_MOUSE  (Q91UZ5), IMPA2_RAT
IMPP_MESCR   (O49071), INM1_YEAST   (P38710), INM2_YEAST
INPP_BOVIN   (P21327), INPP_HUMAN   (P49441), INPP_MOUSE
QAX_NEUCR    (P11634), QUTG_EMENI   (P25416), SUHB_AQUAE
SUHB_ARCFU   (O30298), SUHB_BACSU   (Q45499), SUHB_CAUCR
SUHB_ECO57   (P0ADG6), SUHB_ECOL6   (P0ADG5), SUHB_ECOLI
SUHB_HAEIN   (P44333), SUHB_METJA   (Q57573), SUHB_METTH
SUHB_MYCBO   (P65166), SUHB_MYCLE   (P46813), SUHB_MYCTU
SUHB_NEIMA   (Q9JU03), SUHB_NEIMB   (Q9JZ07), SUHB_PASMU
SUHB_PSEAE   (Q9HXI4), SUHB_RHILO   (Q98F59), SUHB_RHIME
SUHB_SALTY   (P58537), SUHB_SYNY3   (P74158), SUHB_THEMA
SUHB_VIBCH   (Q9KTY5), SUHB_XYLFA   (Q9PAM0), SUHB_XYLFT
Y4FL_RHISN   (P55450), YHEB_CHLVI   (P56160)
```

False negative hits (sequences which belong to the set under consid which have not been picked up by the pattern or profile):

```
DPNP3_ARATH  (Q8GY63), SUHB_AERPE   (Q9YAZ7), SUHB_BUCAI
SUHB_BUCAP   (Q8K9P6), SUHB_BUCBP   (Q89AK9)
```

`Potential' hits (partial sequences which belong to the set under cons which are not hit by the pattern or profile because they are partial (fra sequences):

```
YPSS_RHILP   (P10497)
```

Sequences which could potentially belong to the set under considera

```
PPNK_METJA   (Q58327)
```

Retrieve an alignment of UniProtKB/Swiss-Prot true positive hits:

[Clustal format, color, condensed view] [Clustal format, color] [Clustal form [Fasta format]

PDB [Detailed view]	1AWB; 1DK4; 1FLF; 1G0H; 1G0I; 1IMA; 1IMB; 1IMC; 1IMD; 1IME; 1IMF; 1K9Y; 1K9Z; 1KA0; 1KA1; 1LBV; 1LBW; 1LBX; 1LBY; 1LBZ; 1Q00; 1QGX 2HHM;

View entry in original PROSITE format View entry in raw text format (no links) Direct

ScanProsite submission

If you would like to retrieve all the Swiss-Prot entries referenced in the DR lines of this entry (with the exception of false positive hits) , you can enter a file name. These entries will then be saved to a file under this name in the directory `outgoing` of the ExPASy anonymous ftp server, from where you can download it. (Please note that this temporary file will only be kept for 1 week.)

File name:

Format: ● Swiss-Prot ○ Fasta

[Reset] or [Create file]

ExPASy Home page **Site Map** **Search ExPASy** **Contact us** **PROSITE**

Hosted by SIB Switzerland Mirror sites: Australia Brazil Canada China Korea